Tests psicológicos y evaluación

Lewis R. Aiken

PEARSON
Prentice Hall

Undécima edición

UNDÉCIMA EDICIÓN

TESTS PSICOLÓGICOS Y EVALUACIÓN

LEWIS R. AIKEN
Pepperdine University

TRADUCCIÓN:
María Elena Ortiz Salinas
Universidad Nacional Autónoma de México

Gabriela Montes de Oca Vega
Traductora profesional

REVISIÓN TÉCNICA:
Rubén W. Varela Domínguez
Universidad Nacional Autónoma de México
Instituto Mexicano de Evaluación y Consejería

PEARSON Educación ®

México • Argentina • Brasil • Colombia • Costa Rica • Chile • Ecuador
España • Guatemala • Panamá • Perú • Puerto Rico • Uruguay • Venezuela

Datos de catalogación bibliográfica

AIKEN, LEWIS R.

Tests psicológicos y evaluación. Undécima edición

PEARSON EDUCACIÓN, México, 2003

ISBN: 970-26-0431-1
Área: Universitarios

Formato: 18.5 × 23.5 cm Páginas: 544

Edición en español
Editor: Leticia Gaona Figueroa
e-mail: leticia.gaona@pearsoned.com
Supervisor de desarrollo: Diana Karen Montaño González
Supervisor de producción: José D. Hernández Garduño

Edición en inglés
Executive Editor: *Carolyn Merrill*
Editorial Assistant: *Kate Edwards*
Marketing Manager: *Wendy Gordon*
Editorial Production Service: *Whitney Acres Editorial*
Manufacturing Buyer: *JoAnne Sweeney*
Cover Administrator: *Linda Knowles*

UNDÉCIMA EDICIÓN, 2003

D.R. © 2003 por Pearson Educación de México, S.A. de C.V.
Atlacomulco 500-5to. piso
Industrial Atoto
53519 Naucalpan de Juárez, Edo. de México
E-mail: editorial.universidades@pearsoned.com

Cámara Nacional de la Industria Editorial Mexicana
Reg. Núm. 1031

ISBN 970-26-0431-1

Impreso en México. *Printed in Mexico.*

1 2 3 4 5 6 7 8 9 0 - 06 05 04 03

Cualquier cosa que existe, existe en alguna cantidad. (Thorndike, 1918)

Cualquier cosa que existe en cantidad, puede medirse. (McCall, 1939)

CONTENIDO

PREFACIO

Durante muchos años, los tests y la evaluación en psicología han sido objeto de crítica constante. En repetidas ocasiones se ha atacado el uso de los tests estandarizados, especialmente en contextos educativos y laborales. Ha habido numerosas demandas legales y juicios en los tribunales relacionados con las pruebas psicológicas, por lo que algunos estados de la Unión Americana han instaurado leyes sobre el uso y la reglamentación de los tests. Aunque puede ser justo criticar los métodos para evaluar a las personas y sus actividades, es indiscutible la necesidad de dichos métodos para evaluar, diagnosticar y predecir el comportamiento de los individuos en un mundo con una población de más de seis mil millones de personas. A pesar de las críticas provenientes tanto de profesionales de la psicología y la pedagogía como de otros ámbitos, la evaluación psicológica ha continuado expandiéndose y diversificándose. Como testimonio del dinamismo de los tests y la evaluación en psicología, se encuentran instrumentos nuevos, inventarios y escalas, aunados a los avances metodológicos en cuanto a la elaboración, aplicación, calificación e interpretación de instrumentos psicométricos. Son muchos los factores que han contribuido a este desarrollo, incluyendo la expansión de servicios y las oportunidades sociales hacia un segmento mayor de la población, siempre creciente, así como la necesidad de contar con métodos más efectivos para seleccionar, diagnosticar y ubicar a las personas en contextos laborales, educativos y clínicos.

El desarrollo de los tests psicológicos durante las últimas décadas se ha facilitado por el progreso en el diseño y la programación computacional de alta velocidad. Desde que las computadoras empezaron a estar disponibles comercialmente a mediados de la década de 1950, se han usado para calificar tests y analizar el desempeño tanto de individuos como de grupos. A partir de entonces las computadoras también se han utilizado para aplicar tests y otros instrumentos de evaluación, así como para interpretar sus resultados. Como consecuencia, los tests y otros dispositivos psicométricos literalmente han reestructurado el campo de la evaluación psicológica, y sin duda continuarán haciéndolo en la medida en que lleguen a estar disponibles tecnologías y procedimientos más complejos.

El aumento de la atención del público y de los profesionistas hacia la utilidad y las limitaciones de los tests ha fomentado el deseo de que se incremente el cuidado con que se diseñan y distribuyen tanto los propios tests como otros materiales de evaluación similares. Asimismo, cada vez resulta más obvia y urgente la necesidad de una mejor capacitación entre los usuarios de los tests, y de una mayor conciencia del público y de los profesionales acerca de las consecuencias personales y sociales de las pruebas psicológicas en contextos educativos, clínicos, laborales y empresariales. Los especialistas en psicometría, y otros expertos en tests y en la aplicación de pruebas, se preocupan porque estos instrumentos se diseñen y empleen no sólo prestando atención a sus características técnicas, sino también considerando las necesidades y los derechos de los individuos y de la sociedad en su conjunto. Estos asuntos se abordan en numerosas publicaciones de organizaciones profesionales, tales como la American Psychological Association, la American Educational Research Association, la American Personnel and Guidance Association y el National Council in Measurement on Education.

Consecuente con estas preocupaciones y propósitos, el principal objetivo de este libro de texto es, como lo ha sido desde que se publicó la primera edición hace más de 30 años, mejorar el conocimiento, la comprensión y la práctica de las personas que diseñan tests, los aplican, los

resuelven, los califican, interpretan los resultados y toman decisiones con base en los datos así obtenidos. Al igual que sus predecesoras, la undécima edición está diseñada sobre todo como un libro de texto para estudiantes universitarios. Es adecuado para cursos de un semestre sobre tests y evaluación en un nivel propedéutico o de principiantes en psicología, pedagogía y áreas afines. También puede ser de utilidad para psicólogos y otros profesionales que diseñan y aplican instrumentos de evaluación, e interpretan y aplican los resultados.

Al escribir este libro he intentado abarcar por completo la materia sin llegar a ser exhaustivo, de modo que los instructores que lo adopten descubrirán que no han sido reemplazados por el texto. Éste presenta muchas oportunidades para que el instructor trabaje seleccionando e interpretando, así como reelaborando o ampliando, la información contenida. El Resumen que viene al final de cada capítulo proporciona un panorama y una reseña del material visto en el capítulo, y la sección de Preguntas y Actividades amplía y complementa la información.

La estructura básica de la undécima edición de *Tests Psicológicos y Evaluación* es muy similar a la de las ediciones previas. Los profesores que estén familiarizados con cualquiera de ellas se encontrarán en un territorio conocido que ha cambiado aquí y allá, pero no de manera radical. Algo que los usuarios de ediciones anteriores advertirán de inmediato es que hay más capítulos (18) en esta edición. La estructura de los primeros cinco capítulos es muy similar a la anterior, pero el material del resto del libro se ha redistribuido. Los capítulos 6 a 9 y algunas partes de los capítulos 13 y 14 de la décima edición se han convertido en seis capítulos (6 a 11) en esta nueva edición. El material del capítulo 11 de la décima edición se ha distribuido en tres capítulos (14, 15 y 16) en la actual, y el material que antes se encontraba en el capítulo 12 se ha distribuido ahora en los capítulos 17 y 18. El aumento de capítulos no se debe tanto a que se haya añadido material nuevo, aunque así ocurrió en cierta medida, sino más bien a que se han dividido los anteriores capítulos en otros más breves y se ha incorporado a los capítulos del 6 al 18 material relevante de los capítulos 13 y 14 anteriores. El autor confía en que esta redistribución tenga sentido y facilite el estudio y la comprensión de la información sobre aptitudes cognoscitivas de los capítulos 6 a 11 y el material sobre personalidad, intereses, actitudes y conceptos relacionados, de los capítulos 12 a 18.

En años recientes han ocurrido varios cambios notables, si bien no revolucionarios, en la evaluación psicológica y pedagógica, y se les ha prestado la atención apropiada en este libro. En estos cambios se incluyen revisiones del contenido y el formato de los exámenes de admisión universitarios, las nuevas ediciones de varias pruebas y un interés renovado por la "política de los tests de inteligencia". Se ha dado mayor atención a los tests de adaptación, a la teoría de la respuesta a los ítems, al uso de microcomputadoras en tests psicológicos, a pruebas neuropsicológicas y de desarrollo y a aplicaciones de pruebas en diversos contextos. Para contribuir a lograr el objetivo de introducir los tests psicológicos y la evaluación como un campo de estudio interesante e importante para los estudiantes que planean ingresar a alguno de los diversos campos profesionales en que se diseñan y/o emplean evaluaciones psicológicas, se ha puesto todavía más énfasis en la aplicación de pruebas en ambientes educativos-escolares, clínicos-consultivos e industriales-empresariales.

Se encuentra disponible, en su undécima edición, el libro *Instructor´s Manual to Accompany Psychological Testing and Assessment* (Editorial Allyn & Bacon, Pearson Education). También podría interesarle a los profesores saber que a través del autor aún están disponibles los disquetes de varias docenas de programas de computación que complementan los cursos sobre pruebas psicológicas y educativas. Puede enviar su solicitud, junto con un disquete formateado en sistema DOS y un sobre con estampillas, al doctor Lewis R. Aiken, 3300 Blue Ridge Court, Thousand Oaks, CA 91362. Por último, es posible comprar una *Study Guide* para el texto po-

niéndose en contacto con el autor a la dirección mencionada o en la dirección de correo electrónico laiken@prodigy.net.

He recorrido ya un largo camino con este libro, y el viaje casi ha concluido. Agradezco a todos los estudiantes y colegas que han trabajado con las diez ediciones anteriores y han proporcionado atinadas críticas y sugerencias, así como a los reseñadores de la undécima edición: Angela Hazel, de Rochester College; William Mahler, de Concordia College, y William Warley, de Shorter College. También deseo expresar mi agradecimiento por los esfuerzos incansables y la experiencia de la productora editorial Faye Whitney-Lussier y el editor, William Thomas. Espero sinceramente que los resultados de su trabajo y del mío se manifiesten en el producto terminado. Serán bien recibidos y se agradecerá todo tipo de comentarios y sugerencias para mejorar este libro.

Lewis R. Aiken

CAPÍTULO UNO

TEMAS HISTÓRICOS Y PROFESIONALES

Cualquiera que haya asistido a la escuela básica o a la universidad, ingresado al servicio militar o bien solicitado algún empleo durante el último medio siglo, sin duda ha completado una o más pruebas. En todo el mundo, las pruebas han llegado a tener una gran influencia en la vida y la carrera de las personas. Sin embargo, los instrumentos de evaluación psicológica no se limitan a pruebas publicadas. Se dispone de muchas pruebas inéditas, además de cuestionarios, inventarios, escalas de medición y listas de opción múltiple, tanto publicadas como inéditas.

Siempre que se requiera de información para tomar decisiones con respecto a la gente, o para ayudarla a elegir el rumbo de sus actos relativos a una futura situación educativa o laboral, posiblemente se use algún tipo de instrumento de evaluación. En escuelas, clínicas psicológicas, la industria y el servicio militar y civil, se utilizan ampliamente exámenes y otros instrumentos afines para propósitos de evaluación diagnóstica, selección, asignación y promoción. Además de sus aplicaciones en la toma de decisiones prácticas, las pruebas se usan en forma extensa en la investigación.

Considerando sus múltiples funciones, no es de sorprender que las pruebas, por sí mismas, se hayan convertido en un gran negocio. De acuerdo con la Association of American Publishers, en el año 2000, el total de ventas en Estados Unidos por pruebas estandarizadas aplicadas tan sólo en los grados K-12* (en México equivale al tercer año de bachillerato), se calculaba en 234.1 millones de dólares, una cifra que aumenta en aproximadamente 7% cada año. Hay organizaciones comerciales, como las que figuran en la lista del apéndice C, que se especializan en publicar y distribuir pruebas y otros instrumentos psicométricos para evaluar las aptitudes, personalidades, los intereses y otras características de personas de todas las edades en distintas circunstancias. Las organizaciones profesionales incluidas en el apéndice D se ocupan de lo concerniente al uso adecuado de las pruebas aplicadas con diversos propósitos prácticos y de investigación.

PERSPECTIVA HISTÓRICA

Desde el principio de la historia humana se ha reconocido que las personas difieren en cuanto a sus aptitudes cognoscitivas, características de personalidad y comportamiento, y que estas diferencias pueden evaluarse en cierta forma. Hace casi 2,500 años, Platón y Aristóteles escribieron acerca de las diferencias individuales, e incluso ya tenían como antecesores de esta actividad a los antiguos chinos (Bowman, 1989; Doyle, 1974). Desde la remota fecha del año 2200 a. C., el entonces emperador chino instituyó un sistema de exámenes en el servicio civil para determinar si los funcionarios gubernamentales eran aptos para desempeñar sus labores. Este sistema, de

acuerdo con el cual se examinaba a los funcionarios cada tres años para evaluar su destreza en música, tiro con arco, equitación, escritura, aritmética, así como en ritos y ceremonias públicas y privadas, fue continuado por sucesivos gobernantes chinos, quienes incluyeron conocimientos de la ley civil, asuntos militares, agricultura, rentas públicas, geografía, composición y poesía (Green, 1991). Se trataba de exámenes orales, más que escritos, que evaluaban no solamente lo que los examinados respondían, sino también el cómo lo decían. Durante el siglo XIX, los gobiernos británico, francés y alemán diseñaron sus exámenes para el servicio civil tomando el antiguo sistema chino como patrón.

Durante la Edad Media era prácticamente inexistente cualquier preocupación por la individualidad. En la estructura social de la sociedad europea medieval, las actividades de la gente se determinaban en gran medida dependiendo de la clase social en que se naciera. Se permitía poca libertad para la expresión o el desarrollo individuales. No obstante, hacia el siglo XVI, se tornó más progresista, menos doctrinaria y fue desarrollándose la idea de que las personas eran únicas y tenían derecho a afirmar sus dones naturales y a mejorar su posición en la vida. Esta era de Renacimiento, y el subsiguiente periodo de la Ilustración no sólo fueron etapas durante las cuales el interés por el aprendizaje y la creatividad resurgió y fue fomentado, constituyó también un renacimiento del individualismo. El espíritu de la libertad y el valor individual, que florecía gracias al estímulo político y económico que proporcionaban el capitalismo y la democracia, encontró su expresión en el arte, la ciencia, la filosofía y el gobierno. Sin embargo, no fue sino hasta finales del siglo XIX cuando realmente se inició la evolución del estudio científico de las diferencias individuales en cuanto a aptitudes y personalidad.

Medición mental en el siglo XIX

A principios del siglo XIX, los científicos solían considerar las diferencias en cuanto a habilidades sensoriomotrices y mentales sobre todo como un fastidio o una fuente de error. Antes de la invención de instrumentos precisos y automáticos para medir y registrar acontecimientos físicos, la precisión de las mediciones científicas de tiempo, distancia y otras variables físicas dependía en gran medida de las habilidades de percepción motrices de los observadores humanos. La mayoría de estos observadores estaban muy bien capacitados y eran sumamente cuidadosos al realizar mediciones, pero aun así éstas variaban en forma considerable al ser efectuadas por distintas personas o incluso por el mismo observador en ocasiones diferentes. Debido a que la búsqueda de leyes generales en la naturaleza es difícil cuando las mediciones de fenómenos naturales son imprecisas y no confiables, los físicos dirigieron su atención hacia la construcción de instrumentos que fueran más precisos y consistentes que la sola observación humana. Por ejemplo, la invención que realizaron John Harrison y otros de relojes relativamente libres de los errores ocasionados por el movimiento de los buques y los cambios en temperatura y humedad facilitaron la determinación precisa de la longitud y contribuyeron a hacer los viajes en barco menos azarosos (vea Sobel y Andrewes, 1998).

Impulsado por los escritos de Charles Darwin sobre el origen de las especies y por el surgimiento de la psicología científica, el interés por el estudio de las diferencias individuales creció durante la última parte del siglo XIX. Darwin era inglés, pero la psicología de hecho fue bautizada como ciencia en Alemania al final del siglo XIX. Fue entonces cuando Gustav Fechner, Wilhelm Wundt, Hermann Ebbinghaus y otros psicólogos experimentales demostraron que los fenómenos psicológicos podían ser descritos en términos cuantitativos y racionales. Los acontecimientos que ocurrían en Francia y en Estados Unidos también fueron importantes para el desarrollo de las pruebas psicológicas. La investigación de psiquiatras y psicólogos franceses

sobre perturbaciones mentales influyó en el desarrollo de técnicas de evaluación y tests, y el aumento de la atención dedicada a los exámenes en las escuelas estadounidenses dio como resultado el desarrollo de medidas estandarizadas de los logros académicos.

Al igual que en la historia de cualquier disciplina, muchas personas de varios países desempeñaron papeles significativos en la fase pionera de la medición mental. Especial importancia a fines del siglo XIX tuvieron Francis Galton, J. McKeen Cattell y Alfred Binet. Francis Galton (figura 1.1.), primo del naturalista Charles Darwin, fue un caballero inglés que se interesó en las bases hereditarias de la inteligencia y en la medición de las habilidades humanas. Galton dedicó su atención en particular a la herencia del talento, pero también elaboró una serie de pruebas sensoriomotrices y diseñó varias técnicas para investigar las diferencias individuales en cuanto a aptitudes y temperamento. Usando estas pruebas sencillas, Galton realizó mediciones con más de nueve mil personas, cuyas edades iban de los 5 a los 80 años. Entre sus contribuciones metodológicas figura la técnica de *co-relaciones*, que sigue siendo un método popular para analizar calificaciones de pruebas.

James McKeen Cattell fue un estadounidense que, al regresar de Alemania tras haber obtenido un doctorado en psicología experimental en la Universidad de Leipzig con la tutoría de Wilhelm Wundt, permaneció un tiempo en Inglaterra donde entró en conocimiento de los métodos y pruebas de Galton mientras fungía como su asistente. Más tarde, en la Universidad de Columbia, Cattell intentó relacionar las calificaciones de las mediciones de tiempo de reacción y discriminación sensorial con las calificaciones escolares. Clark Wissler y otros investigadores descubrieron que las relaciones, o correlaciones, entre el desempeño en las pruebas y el logro académico eran muy bajas. Tocó a otro psicólogo francés, Alfred Binet, construir la primera prueba mental que contribuyó en forma significativa a la predicción del aprovechamiento académico.

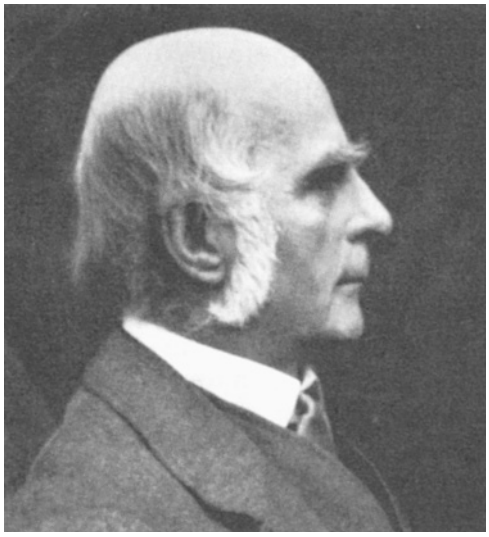


FIGURA 1.1 Francis Galton. El llamado "padre de la psicología individual". Galton fue pionero en el estudio de la inteligencia y sus orígenes.

Las pruebas a principios del siglo xx

En 1904 el ministro de educación pública en París, Francia, comisionó a Alfred Binet (figura 1.2) y a su socio, el doctor Théodore Simon, para que elaboraran un procedimiento de identificación de niños que al parecer fueran incapaces de sacar el provecho suficiente en las aulas escolares normales. Para este propósito, Binet y Simon construyeron una prueba, para ser administrada individualmente, que consistía en 30 problemas dispuestos en orden creciente de dificultad. Los problemas de esta primera *prueba de inteligencia* práctica, que se publicó por primera vez en 1905, pusieron énfasis en la habilidad para juzgar, comprender y razonar. En 1908 se publicó esta prueba revisada, conteniendo entonces una gran cantidad de subpruebas clasificadas por niveles de edad, de los 3 a los 13 años. Al calificar la revisión de 1908 de la Escala de Inteligencia de Binet-Simon, se introdujo el concepto de *edad mental* como una forma de cuantificar el desempeño general de una persona en la prueba. Una revisión adicional de la escala de Binet-Simon, publicada después de la muerte prematura de Binet en 1911, amplió la prueba hasta la edad adulta.

Otros pioneros en pruebas y evaluaciones psicológicas fueron Charles Spearman en teoría de los tests, Edward Thorndike en pruebas de aprovechamiento, Lewis Terman en pruebas de inteligencia, Robert Woodworth y Hermann Rorschach en pruebas de personalidad, y E. K. Strong hijo en mediciones de interés. El trabajo de Arthur Otis con pruebas de inteligencia administradas colectivamente condujo directamente a la elaboración de los Exámenes Alfa y Beta del ejército por parte de un comité de psicólogos durante la Primera Guerra Mundial. Cada uno de estos tests, el Alfa para gente que sabia leer y el Beta para analfabetos, se aplicaba en forma colectiva para medir las habilidades mentales de miles de soldados estadounidenses durante y después de la guerra.

Muchos individuos han contribuido a enriquecer la teoría y la práctica de las pruebas psicológicas y educativas desde la Primera Guerra Mundial. Los nombres de gran parte de ellos se incluyen en la tabla 1.1 y todavía aparecen en los nombres de las pruebas y como referencia a

FIGURA 1.2 Alfred Binet. Con Théodore Simon, en 1905 Binet elaboró la primera prueba de inteligencia práctica.

(Reimpreso con autorización de Culver Pictures, Inc.)

TABLA 1.1 Eventos selectos en la historia de la evaluación psicológica y educativa

1845	Publicación de los primeros exámenes usados por el Comité Escolar de Boston bajo la dirección del educador Horace Mann.
1864	George Fischer, director de escuela inglés, elabora una serie de escalas consistentes en una muestra de preguntas y respuestas como guías para evaluar las respuestas de los estudiantes a preguntas de pruebas de ensayo.
1869	El estudio científico de las diferencias individuales se inicia con la publicación de *Classification of Men According to Their Natural Gifts* (Clasificación de los hombres de acuerdo con sus dones naturales), de Francis Galton.
1882	Emil Kraepelin emplea técnicas de asociación de palabras para estudiar la esquizofrenia.
1884	Francis Galton abre el Laboratorio de Antropometría para la Exposición de Salud Internacional en Londres.
1888	J. M. Cattell abre un laboratorio de pruebas en la Universidad de Pensilvania.
1893	Joseph Jastrow presenta pruebas sensoriomotrices en la Exposición de Columbia en Chicago.
1897	J. M. Rice publica los descubrimientos de su investigación sobre las habilidades ortográficas de los escolares estadounidenses.
1904	Charles Spearman describe su teoría de dos factores sobre aptitudes mentales. Se publica el primer libro de texto importante sobre medición educativa: *Introduction to the Theory of Mental and Social Measurement* (Introducción a la teoría de la medición mental y social), de E. L. Thorndike.
1905	Se publica la primera edición de la Escala de Inteligencia de Binet-Simon.
1908	Se publica la revisión de la Escala de Inteligencia de Binet-Simon.
1908–1909	J. C. Stone y S. A. Courtis publican las pruebas objetivas de aritmética.
1910	Carl Jung elabora una lista estandarizada de estímulos de asociación de palabras para analizar complejos mentales y recopila normas relacionadas.
1908–1914	E. L. Thorndike elabora pruebas estandarizadas de aritmética, caligrafía, lenguaje y ortografía, incluyendo la *Scale for Handwriting of Children* (Escala de caligrafía para niños, 1910).
1914	Arthur Otis elabora la primera prueba de inteligencia colectiva de grupo, basada en la Revisión Stanford de Terman de la Escala de Inteligencia Binet-Simon.
1916	Lewis Terman publica la Escala de Inteligencia de Stanford-Binet.
1917	Los Exámenes Alfa y Beta del ejército, los primeros tests de inteligencia colectivos son elaborados y administrados a los reclutas estadounidenses.
1926	Se aplica por primera vez la Prueba de Aptitud Académica (SAT, por sus siglas en inglés) para evaluar a los aspirantes a ingresar en la universidad.
1927	Se publica la primera edición del Formulario de Intereses Vocacionales para Varones, de Strong, así como las Pruebas de Inteligencia de Kuhlmann-Anderson.
1936	Los Exámenes de Registro de Graduados (GRE, por sus siglas en inglés) se usan por primera vez para seleccionar a los aspirantes a ingresar a la escuela de posgrado.
1937	Se publica la revisión de la Escala de Inteligencia de Stanford-Binet.
1938	Henry Murray Publica *Explorations in Personality* (Exploraciones sobre personalidad). Buros publica el primer *Mental Measurements Yearbook* (Anuario de mediciones mentales).
1939	Se publica la Escala de Inteligencia de Wechsler-Bellevue.
1942	Se publica el Inventario Multifásico de Personalidad de Minnesota.
1949	Publicación de la Escala de Inteligencia de Wechsler para Niños.
1960	Se publica la Forma L-M de la Escala de Inteligencia de Stanford-Binet.
1970–2002	Uso creciente de las computadoras para diseñar, administrar, calificar, analizar e interpretar pruebas.

(continúa)

TABLA 1.1 Continuación

1971	Resolución de la Corte Federal de Estados Unidos para que las pruebas empleadas en la selección de personal estén relacionadas con los puestos (*Griggs* versus *Duke Power*).
1980–2002	Elaboración de la teoría de respuesta.
1981	Se publica una revisión de la Escala de Inteligencia de Wechsler para Adultos.
1985	Se publican los *Standards for Educational and Psychological Testing* (Normas para la evaluación pedagógica y psicológica).
1989	Se publican el MMPI-II y la Escala de Inteligencia para Nivel Preescolar de Wechsler.
1990	Se publica la Escala de Inteligencia para Niños de Wechsler-III.
1997	Aparece la tercera edición de la Escala de Inteligencia para Adultos de Wechsler (WAIS-III).
1998	Se publica la decimotercera edición del *The Mental Measurements Yearbook*.
1999	Se publica *Tests in Print V* y una revisión de los *Standards for Educational and Psychological Testing*.

técnicas, procedimientos y otros adelantos en los que han contribuido. Entre estos progresos se encuentran el perfeccionamiento de la metodología estadística, avances tecnológicos en la preparación y calificación de pruebas y el análisis de resultados en las evaluaciones.

LOS TESTS COMO UNA PROFESIÓN

El campo de aplicación de los tests psicológicos ha crecido rápidamente desde la década de 1920 y en la actualidad se producen y distribuyen comercialmente cientos de estas pruebas. Después de la Segunda Guerra Mundial, las pruebas estandarizadas, en particular las orientadas a evaluar los aprovechamientos académicos, se expandieron por todo el mundo. Muchas pruebas de aptitud y personalidad elaboradas en Estados Unidos se tradujeron del inglés a otras lenguas. Además de las pruebas estandarizadas ya publicadas, pudo disponerse de cientos de materiales de evaluación inéditos. Dichos instrumentos, que se citan en revistas y libros especializados, se han usado en todo el mundo.

Fuentes de información

La información concerniente a tests psicológicos y otros instrumentos de evaluación puede encontrarse en páginas Web y en los catálogos de las empresas que los distribuyen (vea el apéndice C). Muchas de estas compañías publican gran cantidad de catálogos de pruebas. Por ejemplo, la Psychological Corporation tiene catálogos distintos de acuerdo con las áreas de evaluación psicológica, terapia ocupacional y física, habla y lenguaje, y negocios/industria/gobierno. La empresa Pro.ed también cuenta con catálogos por separado para productos como tests psicológicos; educación especial, rehabilitación, trastornos del desarrollo y en superdotados; primera infancia, y habla, lenguaje y audición. En los manuales adjuntos se incluyen más detalles sobre cada una de las pruebas.

También se han publicado varios libros de consulta que abordan el tema de las pruebas. Dos fuentes importantes son: *Tests in Print V* (Murphy, Impara y Plake, 1999) y *Tests* (Maddox, 1997), las cuales proporcionan información descriptiva sobre cientos de pruebas disponibles comercialmente. Otra fuente importante es *The Mental Measurements Yearbook* (Impara y Plake, 1998 y ediciones anteriores), cuyas trece ediciones contienen descripciones y revisiones de

pruebas. También se incluyen revisiones de pruebas en *Test Critiques* (Keyser y Sweetland, 1984-1994).

Tal vez la forma más directa de obtener información sobre pruebas de aplicación común sea consultar ERIC/AE Test Locator, un proyecto conjunto de ERIC Clearinghouse on Assessment and Evaluation de la Universidad Católica de América, la Sección de Biblioteca y de Servicios de Consulta del Educational Testing Service, el Instituto Buros de Mediciones Mentales de la Universidad de Nebraska en Lincoln, el Centro Comprensivo de la Región III de la Universidad George Washington, y los editores Pro-ed test. Es posible entrar en contacto directamente con la página Web de ERIC/AE Test Locator en: www.ericae.net/testcol.htm, www.unl.edu/buros, o bien en www.ets.org. Desde el Test Locator, pueden localizarse seis diferentes archivos: ETS/ERIC Test File, Test Review Locator, BUROS/ERIC Test Publisher Locator, CEEE/ERIC Test Database, los cuales contienen las pruebas que suelen usarse con los estudiantes de LEP, el Reglamento de prácticas de evaluación justas, y Consejos para la selección de pruebas.

Además de las pruebas estandarizadas, en contextos de psicología aplicada se usan muchos cuestionarios y escalas de clasificación (vea Aiken, 1996, 1997). El libro *Measures for Clinical Practice: A Sourcebook* (3ª ed., Corcoran y Fisher, 2000), contiene información descriptiva sobre docenas de instrumentos de este tipo que se utilizan en situaciones de consulta clínica y asesoría.

Para encontrar detalles sobre pruebas y escalas inéditas, también pueden consultarse: *Directory of Unpublished Experimental Mental Measures* (Goldman, Mitchell y Egelson, 1997 y volúmenes anteriores), *A Consumer´s Guide to Tests in Print* (Hammill, Brown y Bryant, 1992), e *Index to Tests Used in Educational Dissertations* (Fabiano, 1989). Para información inédita sobre mediciones de actitudes, se recomienda consultar la serie de volúmenes producidos en el Instituto de Investigación Social de la Universidad de Michigan (Robinson, Shaver y Wrightsman, 1991, 1999 y volúmenes anteriores). En las bibliotecas de muchas universidades está disponible la base de datos HAPI (*Health and Psychosocial Instruments*), que contiene descripciones de más de 15 mil instrumentos psicométricos. Otras bases de datos útiles para obtener información sobre escalas y otros instrumentos psicométricos inéditos son PsycINFO y PsycLIT.

En muchas revistas profesionales se publican versiones de pruebas selectas y revisadas, por ejemplo en: *American Educational Research Journal, Journal of Educational Measurement, Measurement and Evaluation in Counseling and Development, Personnel Psychology* y *Psychoeducational Assessment*. Se incluyen artículos sobre el desarrollo y la evaluación de tests y mediciones psicológicas en publicaciones especializadas como: *Applied Psychological Measurement, Educational and Psychological Measurement, Journal of Clinical Psychology, Psychological Assessment: A Journal of Consulting and Clinical Psychology, Journal of Counseling Psychology* y *Journal of Vocational Behavior.* También pueden encontrarse referencias a fuentes de información sobre pruebas específicas en: *Psychological Abstracts, Education Index* y *Current Index to Journals in Education*. Se han escrito libros enteros sobre pruebas individuales, como el Inventario Multifásico de Personalidad de Minnesota (MMPI), el Test de las Manchas de Tinta de Rorschach y las escalas de inteligencia de Wechsler.

Clasificación de pruebas

Al igual que en otras profesiones, en psicología la evaluación tiene su propio vocabulario especial. El glosario que se incluye al final de este libro contiene definiciones de docenas de términos psicométricos, muchos de los cuales se refieren a tipos de pruebas o a métodos para clasificarlas. Las pruebas se pueden clasificar según su contenido, la forma en que se elaboraron, el parámetro para cuya medición se diseñaron, el propósito de su aplicación, e incluso de acuerdo con

la manera en que se administran, califican e interpretan. Un criterio de clasificación sencillo es la dicotomía entre pruebas *estandarizadas contra no estandarizadas*. Una *prueba estandarizada*, elaborada por profesionales especialistas en desarrollar pruebas y que es administrada a una muestra representativa de personas pertenecientes a la población para la que se diseñó el instrumento, tiene procedimientos establecidos de administración y calificación que son constantes en los distintos examinandos. Así, todos ellos tienen la misma oportunidad de responder los distintos reactivos de acuerdo con sus habilidades.

Por lo general, las pruebas estandarizadas poseen *normas*; esto es, a partir de las puntuaciones crudas obtenidas en la *muestra de estandarización*, se calculan varios tipos de calificaciones transformadas. Las normas sirven como base para interpretar los resultados de las personas que se someten a la prueba después. Todavía más comunes que las pruebas estandarizadas publicadas son los exámenes escolares no estandarizados, los que suelen elaborar los maestros de manera informal.

Las pruebas también se clasifican como *individuales* o *colectivas*. Una *prueba individual*, como la Escala de Inteligencia de Wechsler para Niños, se administra a un examinando en cada ocasión. Una *prueba colectiva,* como la Prueba de Aptitudes Cognoscitivas, puede administrarse simultáneamente a muchos examinandos.

Mientras que la dicotomía de *pruebas individuales contra pruebas colectivas* se refiere a la eficiencia de la administración, la dicotomía de *velocidad contra potencia* corresponde al tiempo límite que se da para resolver una prueba. Una *prueba de velocidad* simple consta de muchos reactivos, pero los límites de tiempo son muy estrictos y casi nadie termina en el lapso asignado. Los límites de tiempo en una *prueba de potencia* son amplios para la mayoría de los examinandos, pero la prueba contiene reactivos más difíciles que los de una prueba de velocidad.

Una tercera dicotomía se presenta en la clasificación: pruebas *objetivas contra no objetivas*, y se refiere al método de calificar una prueba. Una *prueba objetiva* tiene normas de calificación precisas ya establecidas y puede ser calificada por un empleado. Por otra parte, calificar pruebas de ensayo y ciertos tipos de tests de personalidad es muy subjetivo y los resultados pueden variar cuando una misma prueba es calificada por personas distintas.

Las pruebas también pueden clasificarse de acuerdo con el tipo de material o la clase de tarea que se pide a los examinandos. Algunas pruebas sólo contienen reactivos *verbales* o *lingüísticos* (por ejemplo, párrafos de vocabulario o de lectura), mientras que otras consisten en diagramas, rompecabezas u otros materiales *no verbales* o *no lingüísticos*. La distinción entre pruebas verbales y no verbales también se refiere a la forma de la respuesta requerida. Las pruebas que exigen respuestas orales o escritas a menudo reciben el nombre de pruebas *verbales*, mientras las que piden a los examinandos señalar las respuestas correctas, construir algo o manipular materiales de prueba (armar rompecabezas, introducir bloques en agujeros y similares) se denominan pruebas *no verbales* o pruebas *de ejecución.*

Otra clasificación de pruebas amplia, según su contenido o proceso, es en *cognoscitivas contra afectivas*. Las *pruebas cognoscitivas* intentan cuantificar los procesos y productos de la actividad mental y pueden clasificarse como mediciones de rendimiento y aprovechamiento. Una *prueba de rendimiento* evalúa el conocimiento de algún tema u ocupación académica y se centra en el comportamiento pasado del examinando (en lo que ya ha aprendido o logrado). Una *prueba de aprovechamiento* está enfocada al comportamiento futuro, es decir, a lo que la persona es capaz de aprender con la capacitación apropiada. Así, las pruebas de aptitud mecánica y de aptitud para el trabajo de oficina se diseñan para evaluar la habilidad para aprovechar una capacitación adicional en tareas mecánicas y de oficina, respectivamente. Sin embargo, el rendimiento y el aprovechamiento no son entidades separadas; lo que una persona ha alcanzado en el

pasado (rendimiento) suele ser un muy buen indicador de la eficacia con que se desempeñará en el futuro (aprovechamiento). Algunos psicólogos prefieren no usar los términos *rendimiento* y *aprovechamiento* como formas de clasificar pruebas; más bien se refieren a ambos tipos de prueba como medidas de habilidad.

Las *pruebas afectivas* se diseñan para evaluar intereses, actitudes, valores, motivos, rasgos de carácter y otras características de personalidad no cognoscitivas. Para este propósito se han diseñado diversas técnicas, tales como la observación del comportamiento, los inventarios en lápiz y papel y las imágenes proyectivas.

Algunas instituciones y organizaciones que conservan colecciones de pruebas psicológicas y educativas tienen sistemas formales para clasificar estos instrumentos. Uno de los sistemas de clasificación más completos es *The Mental Measurements Yearbook*, donde las pruebas se clasifican en 18 grandes categorías de contenido, las cuales se presentan en la tabla 1.2.

Objetivos y usos de las pruebas

Las pruebas psicológicas y otros instrumentos de evaluación se aplican en un amplio rango de ambientes académicos, clínicos-consultivos, de negocios-industriales, de justicia criminal-forenses, gubernamentales y militares. Los psicólogos de personal, clínicos, consultores, sociales, y muchos otros especialistas dedicados a la investigación o a aplicaciones prácticas en el comportamiento humano, dedican una parte considerable de su tiempo profesional a calificar e interpretar pruebas psicológicas. Las páginas Web de muchas de las compañías que se ocupan de los tests psicológicos y la evaluación se incluyen en el apéndice D.

El objetivo principal de las pruebas psicológicas en la actualidad es el mismo que el prevaleciente en todo el siglo XX: evaluar el comportamiento, las aptitudes cognoscitivas, los rasgos de personalidad y otras características individuales y de grupo, a fin de ayudar a formarse juicios, predicciones y decisiones sobre la gente. De manera más específica, las pruebas se usan para:

1. Seleccionar aspirantes a empleos y programas educativos y de capacitación.
2. Clasificar y colocar a las personas en contextos educativos y laborales.
3. Asesorar y guiar a las personas con propósitos de asesoría educativa, vocacional y personal.
4. Conservar o despedir, promover y rotar estudiantes o empleados en programas educativos, de capacitación y en situaciones laborales.
5. Diagnosticar y prescribir tratamientos psicológicos y físicos en clínicas y hospitales.
6. Evaluar cambios cognoscitivos, intra o interpersonales relativos a programas educativos, psicoterapéuticos y otros de intervención en el comportamiento.
7. Supervisar la investigación sobre cambios en el comportamiento a lo largo del tiempo y evaluar la eficacia de nuevos programas o nuevas técnicas.

TABLA 1.2 Categorías de tests incluidas en *The Thirteenth Mental Measurements Yearbook*

Aprovechamiento	Lenguas extranjeras	Ciencia
Evaluación del comportamiento	Aptitudes de inteligencia y generales	Sensoriomotrices
Desarrollo	Matemáticas	Estudios sociales
Educación	Varios	Habla y audición
Inglés y lenguaje	Neuropsicológicas	Vocacionales
Bellas Artes	Personalidad	
	Lectura	

Además de analizar y describir características individuales, las pruebas pueden utilizarse para evaluar ambientes psicológicos, movimientos sociales y otros acontecimientos psicosociales.

Entre las pruebas que están disponibles comercialmente, no se sabe con exactitud cuántas de cada tipo se usan, en qué situaciones, con qué objetivos ni quién las administra en un año determinado. Sin embargo, puede encontrarse un indicio general de su utilización en los resultados de varias investigaciones (Archer, Mariush, Imhof y Piotrowski, 1991; Butler, Retzlaff y Vanderploeg. 1991; Camara, Nathan y Puente, 2000; Piotrowski y Keller, 1992; Watkins, Campbell y Nieberding, 1994; Watkins, Campbell, Nieberding y Hallmark, 1995). Como es comprensible, los descubrimientos de estos estudios dependen de todo tipo de practicantes y/o investigadores incluidos en la muestra de la investigación, de su orientación teórica y de los objetivos del proceso de evaluación. Las primeras dos secciones de la tabla 1.3 muestran, en orden de rango, las diez pruebas más usadas por los psicólogos clínicos y los neuropsicólogos del estudio de Camara *et al.* (2000). Otro indicador de la popularidad general de un instrumento de evaluación, en particular en investigaciones publicadas, es la cantidad de veces que se menciona en la base de datos PsycINFO. Los diez instrumentos psicométricos empleados con mayor frecuencia entre 1995 y 2001 en contextos clínicos y de asesoría se incluyen en la última sección de la tabla 1.3.

ÉTICA Y NORMAS DE LOS TESTS

El aumento en el uso de pruebas estandarizadas de todo tipo ha dado origen al reconocimiento de la necesidad de ampliar la conciencia pública acerca de las ventajas y limitaciones de los instrumentos de evaluación psicológica y pedagógica, así como las motivaciones y prácticas de quienes las distribuyen y emplean. Una de las preocupaciones constantes de las organizaciones profesionales de psicólogos y educadores es que las pruebas disponibles comercialmente deberían medir efectivamente lo que declaran sus autores, editores y distribuidores. Contribuye al logro de esta meta la edición de 1999 del folleto de normas técnicas *Standards for Educational and Psychological Testing* (AERA, APA y NCME, 1999), que es una modificación del *Standards* de 1985 elaborada por representantes de la Asociación Americana de Investigación Educativa (AERA), la Asociación Americana de Psicología (APA) y el Consejo Nacional sobre Medición en Educación (NCME). Al igual que las entregas anteriores, la edición de 1999 contiene las normas recomendadas para la elaboración y aplicación de pruebas. En ella se abordan con cierto detalle los criterios para evaluarlas, la práctica de su aplicación y los efectos de su uso.

También se ocupan de fomentar la utilización adecuada de pruebas psicológicas y pedagógicas *Guidelines for Computer-based Tests and Interpretations* (American Psychological Association, 1986) y los *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, Inc., 1987).

Preparación de los usuarios de pruebas

La preparación requerida para aplicar, evaluar e interpretar pruebas varía en cierta medida de acuerdo con el tipo de prueba en particular. Las normas de preparación para los usuarios son más estrictas en el caso de pruebas individuales que en pruebas colectivas, y en pruebas de inteligencia y personalidad que en las de rendimiento y aptitudes especiales. Quienquiera que sea el usuario y tenga la preparación que tenga, la responsabilidad ética de garantizar que las pruebas se vendan sólo a personas preparadas corresponde directamente a los editores y distribuidores de

TABLA 1.3 Las diez pruebas usadas con mayor frecuencia por psicólogos clínicos y neuropsicólogos y las diez pruebas más mencionadas en PsycINFO, 1995-2001

PRUEBAS USADAS POR PSICÓLOGOS CLÍNICOS[a]

1. Escala de Inteligencia para Adultos de Wechsler, Revisada (WAIS-R)
2. Inventario Multifásico de Personalidad de Minnesota (MMPI) I y II
3. Escala de Inteligencia para Niños de Wechsler, Revisada (WISC-R y III)
4. Test de las Manchas de Tinta de Rorschach
5. Test Gestáltico Visomotor de Bender
6. Test de Apercepción Temática (TAT)
7. Prueba de Rendimiento de Rango Amplio-R y III
8. Técnica Proyectiva Casa-Árbol-Persona
9. Escala de Memoria de Wechsler, Revisada
10. Inventario de Depresión de Beck, Inventario Multiaxial Clínico de Millon

PRUEBAS USADAS POR NEUROPSICÓLOGOS[a]

1. Inventario Multifásico de Personalidad de Minnesota (MMPI) I y II
2. Escala de Inteligencia para Adultos de Wechsler, Revisada (WAIS-R)
3. Escala de Memoria de Wechsler, Revisada
4. Test de Trazar un Camino A y B
5. Prueba FAS de Fluidez de Palabra
6. Batería de Pruebas Neuropsicológicas de Halstead-Reitan
7. Prueba de Memoria de Boston
8. Prueba de Categoría
9. Prueba de Rendimiento de Rango Amplio-R y III

PRUEBAS MENCIONADAS EN PSYCINFO

1. Escala de Inteligencia para Adultos de Wechsler, Revisada (WAIS-R)
2. Inventario Multifásico de Personalidad de Minnesota (MMPI) I y II
3. Test de las Manchas de Tinta de Rorschach
4. Escala de Inteligencia para Niños de Wechsler, Revisada (WISC-R y III)
5. Indicador Tipológico de Myers-Briggs
6. Inventario de Depresión de Beck
7. Inventario Multiaxial Clínico de Millon
8. Test de Apercepción Temática
9. Lista de Verificación de Conducta para Niños
10. Escala de Memoria de Wechsler, Revisada

[a]Con base en datos proporcionados por Camara, Nathan y Puente, 2000.

las pruebas. Estas organizaciones deben encargarse de explicar y establecer la preparación necesaria para aplicar e interpretar pruebas específicas.

Los prestigiados editores comerciales de pruebas solicitan que los compradores cumplan con ciertos requisitos, dependiendo del carácter de la prueba y/o del grado de preparación necesario para aplicarla. Los Sistemas de Orientación Americanos (AGS) y los de The Psychological

Corporation, así como otras organizaciones comerciales, adoptan una política de preparación de usuarios de tres niveles (A, B y C). AGS define estos tres niveles de la siguiente manera:

> Nivel A: El usuario ha terminado al menos un curso sobre medición, guía o una disciplina similar adecuada, o bien cuenta con la experiencia equivalente supervisada en aplicación e interpretación.
>
> Nivel B: El usuario ha concluido una capacitación graduada sobre medición, guía, evaluación psicológica individual o métodos de valoración especial adecuados para una prueba en particular.
>
> Nivel C: El usuario ha terminado con reconocimiento un programa de capacitación con trabajo apropiado y ha supervisado la experiencia práctica en la administración e interpretación de instrumentos de evaluación clínica.

El formato 1.1, que ha sido adoptado por la AGS para determinar si los compradores de pruebas individuales cumplen los requisitos de cada uno de estos tres niveles, se basa en la investigación realizada por el Grupo de Preparación del Usuario de Pruebas (Moreland, Eyde, Robertson, Primoff y Most, 1995). Esta investigación utilizó métodos de análisis de empleos para describir la capacidad de 86 usuarios de pruebas y 7 factores relacionados con el mal uso de éstas. Los resultados condujeron a los investigadores a concluir que el uso profesional, sano, de pruebas significa que todos los usuarios deben:

1. Mantener la seguridad de los materiales de evaluación antes y después de aplicar las pruebas.
2. Evitar etiquetar a los individuos con base en el resultado de una sola prueba.
3. Respetar estrictamente la ley de derechos de autor y en ninguna circunstancia fotocopiar o reproducir cuadernillos de preguntas, hojas de respuestas, libros de texto ni manuales.
4. Aplicar y calificar las pruebas exactamente en la forma que lo especifica el manual.
5. Entregar los resultados sólo a las personas autorizadas y conforme a los principios aceptados de interpretación de pruebas (Moreland *et al*., p. 23).

Códigos de ética

El uso ético de pruebas puede controlarse en cierta medida mediante un código de ética al que se suscriben los profesionales en aplicación de pruebas y los editores. La Asociación Americana de Psicología (APA), la Asociación Americana de Personal y Asesoría (APGA) y el Consejo Nacional de Medición en Educación (NCME) tienen códigos de ética correspondientes a la aplicación de pruebas y a la prestación de otros servicios psicológicos. Los códigos de ética de la APA, la APGA y el NCME abarcan muchos de los temas de la aplicación, normalización, confiabilidad y validez de las pruebas que se incluyen en los *Standards for Educational and Psychological Testing* (AERA, APA y NCME, 1999). Los tres códigos destacan la importancia de considerar el bienestar del examinando o cliente y de protegerlo del mal uso de los instrumentos de evaluación.

Con respecto a la evaluación y el diagnóstico, en "Ethical Principles of Psychologists and Code of Conduct" (American Psychological Association, 1992, edición corregida en proceso, Web URL http://www.apa.org/monitor/feb01/ethicscode.html) se subraya que la evaluación y el diagnóstico deberían ser realizados sólo en un contexto profesional y por parte de especialistas capacitados y competentes en las pruebas adecuadas. También se pone énfasis en (1) la aplicación de procedimientos científicos para diseñar y seleccionar pruebas y técnicas que sean apropiadas para poblaciones específicas; (2) la interpretación juiciosa de los resultados de las pruebas; (3) el uso cuidadoso de las calificaciones de las

FORMATO 1.1 Formato de preparación del usuario de pruebas

En AGS mantenemos un profundo compromiso con las prácticas profesionales en las pruebas estandarizadas. Para ayudar a garantizar el uso seguro de nuestras evaluaciones, requerimos a las personas que adquieren pruebas de AGS por primera vez llenen este formato. Luego, que lean los "Principios del uso eficaz del test" en la sección final de éste, y lo firmen para indicar que aceptan y cumplirán esos principios.

Nombre _____ Puesto _____ Teléfono ( ) _____

Dirección _____ Ciudad _____ Estado _____ Código postal _____

Empresa para la que trabaja _____ Jefe inmediato_____

Dirección de la empresa _____ Ciudad _____ Estado _____ Código postal _____

FAX _____ Correo electrónico _____

AGS ofrece descuentos de 50% en evaluaciones usadas en proyectos de investigación, y de 40% para programas de capacitación universitarios. Si le interesan, comuníquese a AGS por teléfono o por correo. Consulte el índice para los materiales no incluidos.

Títulos profesionales: (marque todos los que correspondan)

❒ Título en: Área _____ Estado _____ Licencia núm. _____

❒ Miembro de la(s) organización(es) profesional(es)

❒ ACA ❒ AERA ❒ APA ❒ ASHA ❒ CEC ❒ NASP ❒ Otras _____

Nivel de preparación: (marque todos los que correspondan)

❒ Licenciatura Año _____ Institución _____ Especialidad _____

❒ Maestría Año _____ Institución _____ Especialidad _____

❒ Doctorado Año _____ Institución _____ Especialidad _____

Cursos (abajo, marque cada curso terminado y encierre en un círculo el nivel que completó).

N = No graduado, G = Graduado, O = Otro (curso especial que haya terminado, como taller, capacitación laboral, etc.)

	(encierre en un círculo)		(encierre en un círculo)	
❒ Pruebas y mediciones básicas	N G O	❒ Uso de pruebas en consultoría	N G O
❒ Estadística descriptiva	N G O	❒ Evaluación de carreras	N G O
❒ Evaluación de inteligencia	N G O	❒ Evaluación neuropsicológica	N G O
❒ Evaluación de habla, audición y lenguaje	N G O	❒ Otros (anótelos abajo)	
❒ Diagnóstico educativo	N G O	_____	N G O
❒ Curso de evaluación en área de	N G O	_____	N G O
especialización: _____	N G O	_____	N G O

Área especial de competencia: (Anote un tipo de evaluación que use habitualmente y que ilustre mejor su habilidad en la aplicación e interpretación de pruebas.) _____

Principios del uso eficaz de pruebas:

El uso válido y profesional de las pruebas educativas y psicológicas implica que todos los usuarios deben:

1. Mantener la seguridad de los materiales de la prueba antes y después de aplicarla.
2. Evitar etiquetar a las personas con base en un único resultado de una prueba.
3. Respetar estrictamente la ley de derechos de autor y en ninguna circunstancia fotocopiar ni reproducir de ningún otro modo las hojas de respuestas, los cuadernillos de pruebas ni los manuales respectivos.
4. Administrar y calificar las pruebas exactamente como lo especifica el manual.
5. Entregar los resultados sólo a las personas autorizadas y de manera acorde con los principios de la interpretación de pruebas.

Su firma indica que acepta y cumplirá los principios descritos.

Firma _____ Fecha _____

(Reproducido con la autorización del Servicio de Guía Americana de Moreland *et al*., 1995.)

pruebas y los servicios de interpretación, y (4) explicaciones claras pero cuidadosas de los hallazgos de la evaluación. También debería hacerse hincapié en la necesidad de mantener la seguridad de las pruebas si éstas han de tener valor.

La simple existencia de pruebas de alta calidad y de un conjunto de normas y principios para sus editores, distribuidores y consumidores no garantiza que éstas se apliquen e interpreten de manera adecuada. Los encargados de aplicar e interpretar las pruebas son responsables directos de su uso apropiado, como lo reconocen cada vez más los psicólogos profesionales. Desafortunadamente, la capacidad y el conocimiento que poseen muchos asesores, médicos clínicos y otros profesionales son inadecuados para aplicar ciertas pruebas. Por lo tanto, es preciso concientizar a quienes aplican pruebas mentales acerca de las limitaciones de su preparación profesional, de la necesidad de obtener más capacitación y de conseguir ayuda de otros profesionales y fuentes de información actualizadas. Asimismo, los examinadores deben ser capaces de formular juicios éticos acertados mostrándose sensibles a las necesidades tanto de los examinandos como de las organizaciones donde trabajan y de la sociedad en su conjunto.

Consentimiento informado y confidencialidad

El revelar de manera inadecuada datos de pruebas, en especial los identificados con el nombre del examinando, es un asunto que constantemente preocupa a los especialistas en evaluación psicológica. El uso creciente de las computadoras y de los bancos de datos relacionados ha incrementado la necesidad de vigilancia para garantizar que los resultados de las pruebas conservadas en archivos electrónicos en particular sean protegidos adecuadamente contra la revelación impropia. A menos que la ley exija otra cosa, se necesita el *consentimiento informado* de quienes se someten a una prueba o de sus representantes legales antes de entregar los resultados identificados con el nombre del examinando a cualquier persona o institución. El consentimiento informado implica que una persona acepta se entregue información privada porque sabe en qué consiste ésta y con quién será compartida. El formato 1.2 es un consentimiento informado que debe leer y firmar el examinando o alguna persona responsable antes de que se lleve a cabo un examen psicológico. Como se señala en este formato, antes de que cualquier prueba u otros procedimientos psicológicos se administren, debe comunicarse al examinando el carácter y los objetivos de la evaluación, por qué se está aplicando, quién tendrá acceso a la información y cómo se usará ésta. Además de los derechos de consentimiento informado y confidencialidad, deberá adjudicarse la "etiqueta menos estigmatizante" al informar sobre la presencia de ciertos síntomas, trastornos y otros problemas psicológicos. Por ejemplo, "incapacitado mentalmente" es a todas luces menos estigmatizante, en el aspecto personal y social, que "débil mental", "idiota" o "retrasado", así como "reacción de adaptación de la adolescencia" lo es menos que "personalidad psicópata".

Desde un punto de vista legal, los datos provenientes de pruebas psicológicas son *comunicaciones privilegiadas* que pueden compartirse con personas ajenas únicamente en casos de absoluta necesidad. En el momento de la administración de la prueba debe avisarse a los examinandos por qué están siendo sometidos a ella, quién tendrá acceso a la información y cómo se utilizará ésta. Después de las pruebas, los examinandos también tienen el derecho de conocer sus resultados y lo que significan. Excepto en circunstancias excepcionales, como cuando una persona es peligrosa para sí misma o para otros, la información de las pruebas es confidencial y no debe revelarse sin el consentimiento informado necesario. Incluso con consentimiento informado, los datos pueden ser *privilegiados*. Esto significa que exceptuando al examinando y, en caso de menores o de personas legalmente incompetentes, alguno o ambos padres o tutor, sólo el abogado del examinando, su médico o psicólogo pueden obtener una copia de la información.

CONSENTIMIENTO INFORMADO PARA UN EXAMEN PSICOLÓGICO

Yo, _____, voluntariamente acepto actuar como participante en un examen psicológico conducido por_____.
He recibido una explicación clara y completa sobre el carácter general y los propósitos del examen y de las razones específicas por las que se me examina. También he sido informado de los tipos de pruebas y demás procedimientos que se aplicarán, así como de la manera en que se utilizarán los resultados.

Me doy cuenta de que quizá no le sea posible al examinador aclararme todos los aspectos del examen mientras éste no haya terminado. También entiendo que puedo poner fin a mi participación en el examen en cualquier momento y sin represalias. Además comprendo que se me informará de los resultados y que éstos no serán entregados a nadie más sin mi autorización. En este momento, solicito que se envíe una copia de los resultados de este examen a:

_____

_____

Firma del examinando	Nombre del examinando en letra de molde
Fecha	Firma del examinador

FORMATO 1.2 Formato para obtener consentimiento informado y conducir un examen psicológico

Las personas legalmente responsables no sólo tienen derecho al acceso a los descubrimientos que los informes de sus propias pruebas ofrezcan, también pueden disponer que se transmitan sus resultados a organismos educativos, clínicos o de asesoría para su uso apropiado. Asimismo, debe hacerse el máximo esfuerzo por mantener la confidencialidad de los resultados de las pruebas y de cualquier información personal. El Acta Familiar de los Derechos Educativos y de Privacía de 1974 establece, por ejemplo, que los resultados de pruebas y otros registros de estudiantes en poder de instituciones educativas que reciben fondos federales pueden ponerse a disposición, en forma identificable por persona, de otros sólo con el consentimiento por escrito del estudiante, de sus padres o de su tutor. Sin embargo, esta acta sí permite a los padres y al personal de la escuela con un "legítimo interés educativo" revisar los registros escolares, al igual que la Ley Pública 94-142 en el caso de niños con discapacidades.

En 1988, el Comité Adjunto de Prácticas de Exámenes publicó un conjunto de derechos y responsabilidades diseñado "para enumerar y esclarecer las expectativas que razonablemente puedan tener quienes se someten a pruebas sobre el proceso de aplicación de éstas, y las expectativas que pueden tener quienes elaboran, aplican y usan las pruebas sobre los que se someten a ellas". La responsabilidad fundamental de someterse a una prueba es asegurarse de comprender los derechos que se tienen y actuar en consecuencia de la manera apropiada (vea la página Web url://www.apa.org/science/jctpweb.html).

RESUMEN

Las raíces de las pruebas psicológicas y la evaluación pueden rastrearse hasta la Grecia y China antiguas, aunque un método concertado, científico, para efectuar la medición de las diferencias individuales en cuanto a aptitudes y personalidad no se estableció sino hasta fines del siglo XIX en Europa y Estados Unidos. El campo de la evaluación psicológica y pedagógica se desarrolló

con rapidez en el siglo XX, y se emplearon ampliamente instrumentos psicométricos de diverso tipo en situaciones educativas, clínicas, de negocios, de gobierno y militares. Estos instrumentos pueden clasificarse en varias formas: estandarizados o no estandarizados, individuales o colectivos, de velocidad o de potencia, objetivos o no objetivos, verbales o no verbales, de lápiz y papel o de ejecución, y cognoscitivos, afectivos o psicomotrices. Los *The Mental Measurements Yearbooks* constituyen la fuente de información más amplia sobre pruebas. En *Test Print V* puede encontrarse una lista descriptiva bastante completa de pruebas, inventarios, escalas, listas de verificación y cuestionarios publicados.

Las pruebas psicológicas y educativas se han convertido en un gran negocio durante las últimas décadas, pero el desarrollo de este campo se ha visto acompañado por debates acerca de la validez y utilidad de las pruebas y sobre la preparación profesional de quienes las aplican e interpretan. La información obtenida de la aplicación de tests psicológicos debe mantenerse confidencial y, salvo algunas excepciones, sólo puede compartirse con otras personas después de haber obtenido el consentimiento por escrito del examinado o de sus tutores o asesores legales.

Con el fin de tener un mejor control que permita evitar el uso inadecuado de las pruebas, la American Psychological Association, la American Educational Research Association, la American Personnel and Guidance Association y el National Council on Measurement in Education han publicado estándares y códigos sobre las prácticas éticas y justas de la aplicación de las pruebas. El cumplimiento de dichos estándares y códigos ayuda asegurar que las pruebas psicológicas y demás instrumentos y procedimientos psicométricos son aplicados por personal calificado de manera tanto sensible como sensata y que los resultados se interpretan y aplican con precisión y consideración.

PREGUNTAS Y ACTIVIDADES

1. Identificar las contribuciones que cada una de las siguientes personas realizó a la evaluación psicológica y educativa: Alfred Binet, J. McKeen Cattell, Francis Galton, Hermann Rorschach, Charles Spearman, Lewis Terman, Edward Thorndike, Robert Woodworth y E. K. Strong, hijo. Para obtener más información, consulte artículos especializados o determinados capítulos en libros que traten acerca de la historia de las pruebas psicológicas y educativas (por ejemplo, French y Hale, 1990; Goldstein y Hersen, 1990; McReynolds, 1986, y Sokal, 1987).

2. ¿Qué procedimientos o instrumentos se usaban en épocas antiguas para evaluar las aptitudes y personalidad de la gente, y cómo se usaban los resultados de aquellas pruebas?

3. Describa y evalúe diversas formas de clasificar las pruebas psicológicas y otros instrumentos de evaluación psicométrica.

4. Examine en alguna biblioteca ejemplares de *The Mental Measurements Yearbooks* (Impara y Plake, 1988 y ediciones anteriores), *Tests* (4ª ed.) (Maddox, 1997), *Test Critiques* (Keyser y Sweetland 1984-1994) y *Tests in Print V* (Murphy, Impara y Plake, 1999). Describa los diversos tipos de información que contienen estas obras de referencia.

5. Se supone que los psicólogos son profesionales que piensan ante todo en el bienestar del público, así como científicos cuya búsqueda de la verdad no permite la explotación de otras personas; entonces, ¿por qué es necesario tener un código de ética explícito que regule la práctica de la psicología en general y de las pruebas psicológicas en particular?

6. ¿De qué manera los conceptos de *consentimiento informado* y *confidencialidad* en las pruebas psicológicas difieren del de *comunicación privilegiada* usado en las leyes y la medicina?

7. Revise el directorio telefónico de una ciudad grande e indague acerca de servicios de evaluación y pruebas educativas. Debe buscar en varias secciones: psicólogos, asesoría, pruebas, exámenes y similares.

8. **(a)** Conéctese a la página Web www.apa.org.
(b) Oprima el botón del *mouse* en la palabra "Students".
(c) En "Topics", oprima sobre Testing.
(d) Explore la página de Testing and Assessment seleccionando los distintos temas resaltados.

CAPÍTULO DOS

DISEÑO Y ELABORACIÓN DE TESTS

La cantidad de esfuerzo invertido en la elaboración de un test psicológico o educativo varía con el tipo de prueba y con los propósitos para los cuales se crea. Es probable que la mayoría de los maestros dedique relativamente poco tiempo a preparar pruebas de ensayo o de respuesta corta para evaluar el progreso de sus alumnos en una unidad de enseñanza. Por otro lado, las pruebas de habilidad y de personalidad diseñadas por especialistas en evaluación psicológica por lo general requieren del esfuerzo de muchos individuos que trabajan por periodos prolongados.

Los procedimientos empleados en la elaboración de una prueba también varían con el tipo de ésta y los propósitos de los usuarios. Preparar un inventario de lápiz y papel, de intereses o de características de la personalidad, implica problemas diferentes a los de construir una prueba de aprovechamiento o de aptitud. De igual modo, los complejos procedimientos seguidos por los diseñadores profesionales de pruebas son poco familiares para la mayoría de los maestros. Cualquiera que sea el tipo de prueba o las metas de los usuarios, se necesita cierto grado de planeación del contenido antes de escribir los reactivos que contendrá el instrumento. La planeación de la prueba deberá incluir definiciones claras de las variables o constructos que van a medirse, descripciones de las personas que van a ser examinadas, las condiciones bajo las cuales se administrará la prueba, e información concerniente a la calificación, interpretación de las puntuaciones y uso que se dará a los resultados.

PLANEACIÓN DE UN TEST

La elaboración de un test requiere la consideración cuidadosa de sus propósitos específicos. Las pruebas cumplen muchas funciones diferentes, y su proceso de elaboración varía en cierto grado de acuerdo con el propósito que se pretenda lograr. Por ejemplo, se siguen procedimientos diferentes al elaborar pruebas de aprovechamiento, de inteligencia, de aptitud especial o un inventario de personalidad. Sin embargo, de manera ideal, la elaboración de una prueba u otro instrumento psicométrico empieza con la definición de las variables o constructos que van a medirse y con el esbozo del contenido propuesto.

Pruebas de observación

La elaboración de una prueba de aptitud para observar a solicitantes de un trabajo particular comienza con un análisis detallado de las actividades que componen ese trabajo. Un análisis de tareas, o *análisis de trabajo*, consiste en especificar los componentes del trabajo de modo que puedan construirse las situaciones de prueba o reactivos más adecuados para predecir el desempeño del empleado. Esas especificaciones pueden incluir *incidentes cruciales*, conductas que son decisivas para el desempeño exitoso o fallido, así como otra información que describa las actividades del trabajo. Dado que la descripción de un trabajo particular por lo general es larga y requiere de-

dicación, la prueba final no medirá todos los aspectos del desempeño del empleado. Tratará sólo con una muestra de los comportamientos más importantes relacionados con el trabajo en cuestión, muestra que en el mejor de los casos debe ser representativa de todas las tareas a realizar.

Pruebas de inteligencia

En el capítulo 7 se describen con detalle los procedimientos empleados por los diseñadores de pruebas de inteligencia, por lo que aquí sólo se hará una breve descripción. Como en la elaboración de cualquier otra prueba, se reúne un conjunto de reactivos que supuestamente miden algún aspecto del constructo "inteligencia". Esos reactivos pueden ser elaborados de acuerdo con una teoría específica de la conducta inteligente o haciendo referencia sólo a los tipos de tareas que la gente muy inteligente puede realizar de manera más efectiva que las personas menos inteligentes. La selección de los reactivos a incluir en la prueba final puede hacerse con base en las relaciones de las respuestas dadas a reactivos con criterios tales como la edad cronológica, así como con las relaciones entre los reactivos de la prueba.

Inventarios y escalas de personalidad

Al elaborar inventarios de personalidad y escalas de calificación se han empleado varios enfoques, algunos basados en el sentido común , otros en las teorías de personalidad y otros más en procedimientos estadísticos. Como se describe en los capítulos 16 y 17, muchos de los instrumentos de evaluación de la personalidad publicados recientemente han sido elaborados combinando enfoques teóricos, racionales y empíricos. Uno o más de estos enfoques pueden emplearse en diferentes etapas del desarrollo del instrumento.

Pruebas de rendimiento

Se ha dedicado más atención a los procedimientos usados para elaborar pruebas de rendimiento académico que a los de otras clases de pruebas. Esto es comprensible cuando nos percatamos de que se aplican más pruebas de rendimiento que todos los otros tipos de pruebas combinados. A pesar del uso generalizado de las pruebas de rendimiento, la mayoría de los profesores, quienes supuestamente están familiarizados con su materia de estudio, no dedica tiempo suficiente a la evaluación del progreso de los estudiantes. Con mucha frecuencia los maestros consideran que los exámenes son algo desagradable adjunto a la enseñanza, en lugar de verlos como parte integral y formativa del proceso educativo. Sin embargo, cuando se usan de manera efectiva, los resultados de los exámenes no se limitan a la sola evaluación y motivación de los estudiantes. También proporcionan información a los maestros, al personal administrativo y a los padres, concerniente a la medida en que se han alcanzado los objetivos educativos específicos. Al proporcionar datos sobre la efectividad del currículo escolar y los procedimientos de enseñanza, las puntuaciones de los exámenes pueden contribuir significativamente en la planificación educativa para estudiantes individuales o grupos, incluso para distritos escolares enteros.

Preguntas para las personas que planifican los instrumentos

Quienes planifican las pruebas de rendimiento de un salón de clases deben empezar por atender con cuidado las siguientes preguntas:

1. ¿Cuáles son los temas y materiales sobre los que se examinará a los estudiantes?
2. ¿Qué tipos de preguntas deben elaborarse?

3. ¿Qué formatos o esquemas de reactivos y pruebas deben utilizarse?
4. ¿Cuándo, dónde y cómo debe administrarse la prueba?
5. ¿Cómo debe calificarse y evaluarse la prueba resuelta?

Las preguntas 1, 2 y 3 se analizan en este capítulo, las preguntas 4 y 5 en el capítulo 3.

Taxonomías de objetivos cognoscitivos

Así como elaborar una prueba de observación para usar en la selección de personal requiere un análisis preliminar del trabajo a desempeñar, la preparación de una prueba para medir objetivos instruccionales específicos es más efectiva cuando las conductas que van a evaluarse se definen claramente al inicio. Desde mediados de la década de 1950 se ha prestado mucha atención a los sistemas formales y estándar de clasificación de los objetivos cognoscitivos, afectivos y psicomotrices de la instrucción. En la tabla 2.1 se presentan las principales categorías de cuatro de esas taxonomías de objetivos cognoscitivos. Las seis principales categorías de la primera taxonomía, la *Taxonomía de objetivos educativos: el dominio cognoscitivo* (Bloom y Krathwohl, 1956), se presentan en orden de la más simple a la más compleja. Esas categorías no son exclusivas, sino más bien progresivamente in-

TABLA 2.1 Compendio ilustrativo de los objetivos cognoscitivos

Bloom y Krathwohl (1956)
- Conocimiento
- Comprensión
- Aplicación
- Análisis
- Síntesis
- Evaluación

Educational Testing Service (1965)
- Memoria
- Comprensión
- Pensamiento

Ebel (1979)
- Comprensión de la terminología (o vocabulario)
- Comprensión del hecho y del principio (o generalización)
- Habilidad para explicar o ilustrar (comprensión de relaciones)
- Habilidad para calcular (problemas numéricos)
- Habilidad para predecir (qué es probable que suceda bajo condiciones especificadas)
- Habilidad para recomendar la acción apropiada (o algunas situaciones de problemas prácticos específicos)
- Habilidad para formular un juicio evaluativo

Gerlach y Sullivan (1967)
- Identificación
- Nominación
- Descripción
- Elaboración
- Ordenamiento
- Demostración

clusivas. Por ejemplo, tanto el Conocimiento (categoría I) como la Comprensión (categoría II) son esenciales para la Aplicación (categoría III) y por ende están incluidas en la tercera categoría. En la tabla 2.2 se presenta una descripción de las categorías registradas en esta taxonomía.

Otra taxonomía presentada en la tabla 2.1, la propuesta por Gerlach y Sullivan (1967), enfatiza la conducta del examinado en la identificación, nominación, descripción, elaboración, ordenamiento o demostración de algo. La *identificación* consiste en indicar qué miembro de un conjunto pertenece a una categoría particular. En la *nominación* debe proporcionarse la etiqueta verbal correcta para un referente o conjunto de referentes. La *descripción* consiste en reportar categorías relevantes de objetos, eventos, propiedades o relaciones. En la *elaboración* se crea un producto de acuerdo con ciertas especificaciones. El *ordenamiento* requiere arreglar en un orden específico dos o más referentes, y la *demostración* consiste en realizar ciertas acciones para cumplir una tarea especificada.

La aplicación de cualquiera de las taxonomías presentadas en la tabla 2.1 debe alentar a la persona que diseña la prueba a ir más allá de los reactivos que miden el reconocimiento simple o la memoria, y a preparar reactivos que midan objetivos educativos de orden superior y demanden reflexión. Los siguientes reactivos, que pueden presentarse en un formato de ensayo o de prueba objetiva, ejemplifican lo anterior:

> ¿Cuál es la fórmula para calcular el error estándar de medición? (*Conocimiento*)
>
> Examine la siguiente gráfica y determine cuántos reactivos deben agregarse a una prueba de 50 reactivos para aumentar su confiabilidad de .60 a .80. (*Comprensión*)

TABLA 2.2 Categorías de la *Taxonomía de objetivos educativos: el dominio cognoscitivo*

I. *Conocimiento* implica el recuerdo de hechos específicos. Los verbos de muestra en los reactivos de conocimiento son *definir, identificar, mencionar* y *nombrar.* Ejemplo de un reactivo de conocimiento es: "Mencione las seis categorías principales de *La taxonomía de objetivos educativos: el dominio cognoscitivo*".

II. *Comprensión* significa entender el significado o propósito de algo. Los verbos de muestra en los reactivos de comprensión son *convertir, explicar* y *resumir*. Ejemplo de un reactivo de comprensión es: "Explique lo que quiere decir el revisor de la prueba cuando dice que ésta no es confiable".

III. *Aplicación* implica utilizar la información y las ideas en situaciones nuevas. Los verbos de muestra en los reactivos de aplicación son *calcular, determinar* y *resolver*. Ejemplo de un reactivo de aplicación es: "Calcule la media y la desviación estándar del siguiente grupo de calificaciones".

IV. *Análisis* es descomponer algo para revelar su estructura y las interrelaciones que hay entre sus partes. Los verbos de muestra en los reactivos de análisis son *analizar, diferenciar* y *relacionar*. Ejemplo de un reactivo de análisis es: "Analice esta unidad instruccional en varias categorías conductuales y de contenido".

V. *Síntesis* es combinar varios elementos o partes en un todo estructural. Los verbos de muestra en los reactivos de síntesis son *diseñar, crear, formular* y *planificar*. Ejemplo de un reactivo de síntesis es: "Diseñe una tabla de especificaciones para una prueba de estadística elemental".

VI. *Evaluación* es formular un juicio basado en el razonamiento. Los verbos de muestra en los reactivos de evaluación son *comparar, criticar, evaluar* y *juzgar*. Ejemplo de un reactivo de evaluación es: "Evalúe el procedimiento usado en la estandarización de esta prueba".

Fuente: Tomado de *Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I: The Cognitive Domain*, por Benjamin S. Bloom *et al*.

Calcule el error estándar de estimación para una prueba que tiene una correlación de .70 con un criterio que tiene desviación estándar de 10. (*Aplicación*)

Distinga entre una prueba de rendimiento para el salón de clase y una prueba estandarizada de rendimiento en términos de lo que mide cada una y la manera en que se utilizan. (*Análisis*)

Formule una teoría que relacione los intereses con la personalidad y mencione la evidencia de investigación apropiada que la apoya. (*Síntesis*)

Evalúe las críticas concernientes al contenido y los usos del SAT. (*Evaluación*)

Objetivos afectivos y psicomotrices

Una función importante de la educación es inculcar en los estudiantes ciertas actitudes, valores y otros estados afectivos. No existe un método que sea completamente satisfactorio para clasificar los objetivos afectivos de la instrucción, pero se ha propuesto toda una serie de sistemas de clasificación. Un ejemplo es la *Taxonomía de objetivos educativos: dominio afectivo* (Krathwohl, Bloom y Masia, 1964). Las principales categorías de esta taxonomía son: I. Recibir o atender. II. Responder o participar. III. Valorar o creer en el valor de algo. IV. Organizar los valores en un sistema. V. Caracterización mediante un valor o valor complejo. En comparación con su contraparte en el dominio cognoscitivo, esta taxonomía no se ha aplicado con mucha frecuencia.

También se han propuesto taxonomías de objetivos educativos en el dominio psicomotriz (por ejemplo, Harrow, 1972; Nixon y Jewett, 1980; Simpson, 1966). Las seis categorías en la *Taxonomía del dominio psicomotriz* de Harrow, por ejemplo, son: movimientos reflejos, movimientos básicos fundamentales, habilidades perceptuales, habilidades físicas, movimientos hábiles y comunicación no discursiva. Los niveles inferiores de la taxonomía de Nixon y Jewett se interesan en la percepción de los componentes de un movimiento y en los esfuerzos de los examinados por repetirlo o recordarlo después de que se ha demostrado. Los niveles superiores ponen de relieve la creación de movimiento para una situación específica en los deportes, la danza u otras actividades físicas. Al aplicar dicha taxonomía, deben tomarse decisiones con respecto a los pesos numéricos que se asignarán a cada componente del desempeño y a si se harán deducciones por errores, torpezas y falta de pulcritud.

Tabla de especificaciones

La mayoría de los diseñadores de pruebas no se adhiere rígidamente a una taxonomía formal al especificar los objetivos que van a medirse. No obstante, al planificar una prueba es útil construir una tabla de especificaciones de dos vías. En dicha tabla, los objetivos conductuales que van a evaluarse se presentan en los encabezados de renglón y los objetivos de contenido (temáticos) como encabezados de columna. Luego se escriben en el cuerpo (celdas) de la tabla las descripciones de los reactivos específicos que caen bajo los encabezados apropiados de renglón y columna.

Una tabla de especificaciones debe ser razonablemente detallada en términos del conocimiento y las habilidades que se espera demuestren los examinados, pero es importante no enfatizar en exceso un objetivo particular. Por ejemplo, puede ser más sencillo elaborar reactivos que evalúen el conocimiento de términos y hechos que reactivos donde se mida la habilidad de analizar y evaluar, pero en la prueba deben incluirse reactivos de las dos últimas categorías.

La tabla 2.3 es una tabla de especificaciones para una unidad sobre preparación, aplicación y análisis de reactivos de pruebas. Advierta que el número total de reactivos que se dedica

a cada tema aparece entre paréntesis debajo del mismo. Una vez que se ha determinado un conjunto de objetivos para un curso o unidad de instrucción y que se ha preparado el bosquejo temático, pueden elaborarse los reactivos de la prueba para medir el grado en que los estudiantes han alcanzado los objetivos mencionados para cada tema.

Ciertos tipos de reactivos de prueba son más apropiados que otros para medir la obtención de objetivos específicos. Los reactivos de respuesta corta y de completamiento son adecuados para evaluar el conocimiento de la terminología, pero inadecuados para evaluar habilidades cognoscitivas de orden superior. Por esta razón, la tabla de especificaciones para una prueba debe ser inspeccionada con cuidado antes de decidir qué tipo de reactivos y cuántos de cada uno son apropiados. Al planifi-

TABLA 2.3 Especificaciones para una prueba sobre preparación y administración de pruebas

OBJETIVO CONDUCTUAL	**CONTENIDO (TEMA)**				
	Preparación	*Elaboración*	*Aplicación*	*Calificación*	*Análisis de reactivos*
Conocimiento de la terminología	Análisis de trabajo; incidentes críticos; muestra representativa (3 reactivos)	Reactivo de aparejamiento; colectivo en espiral; grupo de respuestas (5 reactivos)	Rapport; efecto de halo (2 reactivos)	Clave de lista; calificación compuesta; calificación con máquina (3 reactivos)	Criterio; consistencia interna; homogeneidad de la prueba (3 reactivos)
Conocimiento de hechos específicos	Categorías en la *Taxonomía de objetivos educativos* (2 reactivos)	Ventajas y desventajas de los reactivos de ensayo y de los reactivos objetivos (4 reactivos)	Factores que afectan el desempeño en la prueba (3 reactivos)	Reglas para calificar pruebas de ensayo y pruebas objetivas (3 reactivos)	Métodos para determinar la validez de los reactivos; propósitos del análisis de reactivos (3 reactivos)
Comprensión	Explicación de los propósitos de hacer plan de prueba (2 reactivos)	0 reactivos	0 reactivos	Efectos de la ponderación de los reactivos sobre la calificación total (1 reactivo)	Explicación de la relación entre *p* y D (1 reactivo)
Aplicación	Especificaciones para una unidad sobre examinación (1 reactivo)	Ejemplos de reactivos de opción múltiple para medir comprensión, aplicación, análisis, síntesis y evaluación (4 reactivos)	Instrucciones para una prueba (2 reactivos)	Corrección para la adivinación; ponderación de confianza; uso del nomograma para calificar los reactivos de reordenamiento (4 reactivos)	Cálculo de los índices de dificultad y discriminación; distribución de las respuestas a los distractores (4 reactivos)
Total	8 reactivos	13 reactivos	7 reactivos	11 reactivos	11 reactivos

car una prueba también es necesario considerar cuestiones prácticas como el costo, el tiempo disponible para la administración, la disposición de los reactivos y las condiciones de la prueba.

PREPARACIÓN DE LOS REACTIVOS DEL TEST

La meta principal de la planificación de la prueba es la preparación de un bosquejo detallado, como una tabla de especificaciones, que sirva como guía al elaborar los reactivos con los que se van a evaluar o predecir ciertos objetivos. Una vez preparada dicha tabla o el bosquejo detallado del contenido de la prueba, el siguiente paso es elaborar los reactivos. Por lo general, es recomendable que, en las pruebas objetivas, inicialmente se prepare alrededor de 20% más de reactivos de los que en realidad se necesiten, de modo que se disponga de una cantidad adecuada de buenos reactivos para la versión final de la prueba. Las organizaciones comerciales de tests, como el Educational Testing Service, emplean como elaboradores de reactivos a personas que poseen un conocimiento profundo de la materia de la prueba y destreza suficiente para la creación de reactivos. Cualquier persona que desee aprender cómo elaborar buenos reactivos puede beneficiarse al inspeccionar una muestra de reactivos de las pruebas publicadas, ya que éstos se encuentran entre los mejores disponibles.

Todos los reactivos representan procedimientos para obtener información acerca de los individuos, pero la cantidad y los tipos de información varían con la naturaleza de las tareas planteadas por diferentes tipos de reactivos. Pedir a los examinados que comparen la Batalla de Bulge con la Batalla de Hastings requiere un tipo de respuesta diferente a la que se obtiene cuando se les pide señalar, de entre una serie de acontecimientos, los que ocurrieron en cada batalla. En el primer reactivo se requieren habilidades de organización e integración complejas, mientras que sólo se necesita memoria de reconocimiento para responder al segundo.

Se han sugerido varios métodos para clasificar los reactivos de acuerdo con el formato o la forma de la respuesta requerida. *Completamiento o llenado* contra *selección, recuerdo* contra *reconocimiento,* y *construcción de respuesta* contra *identificación* son formas de diferenciar entre los reactivos donde se pide a los examinados que escriban o construyan una respuesta y aquellos en que se les pide señalar cuál de varias alternativas es correcta. Otro método popular de clasificación de reactivos es *ensayo* contra *objetivo*, de los cuales se presentan ejemplos en la tabla 2.4. Todos los reactivos de ensayo son del tipo de completamiento o llenado, donde la respuesta del examinado se construye en lugar de ser meramente identificada.

Un reactivo objetivo puede ser del tipo de completamiento o llenado, o de selección, dependiendo de si se pide que los examinados construyan una respuesta o seleccionen la mejor respuesta de entre una lista de alternativas. El rasgo crucial de los reactivos objetivos no es la forma de la respuesta, sino la objetividad con la que pueden calificarse. Dos o más calificadores de un reactivo de ensayo pueden estar en desacuerdo en si una respuesta dada es correcta y en cuántos puntos debería recibir. Sin embargo, salvo que ocurran errores administrativos, los diferentes calificadores de una prueba objetiva asignarán la misma calificación a una determinada prueba.

Reactivos de ensayo

La ventaja principal de los reactivos de ensayo es que pueden medir la habilidad personal para organizar, relacionar y comunicar, conductas que no son fáciles de evaluar con los reactivos objetivos. Las pruebas de ensayo tienen las ventajas de que requieren menos tiempo para su elaboración y reducen la probabilidad de que los examinados respondan en forma correcta a los reactivos por

TABLA 2.4 Ejemplos de varios tipos de reactivos de test

I. Reactivos de ensayo

Instrucciones: Escriba una respuesta de media página para cada uno de los siguientes reactivos.

1. Compare las ventajas y desventajas de los reactivos de ensayo y los reactivos objetivos.
2. Explique las razones para realizar un análisis de reactivos en una prueba para el salón de clases.

II. Reactivos objetivos

A. *Respuesta corta*

Instrucciones: Escriba la(s) palabra(s) apropiada(s) en cada espacio.

1. La única cosa que es objetiva acerca de una prueba objetiva es _____ .
2. ¿Cuál es el primer paso formal en la elaboración de una prueba para predecir el desempeño laboral?_____ _____.

B. *Verdadero-falso*

Instrucciones: Encierre V en un círculo si la afirmación es verdadera; encierre F en un círculo si la afirmación es falsa.

V F 1. El sistema de clasificación de pruebas más global es el de *The Mental Measurements Yearbooks*.

V F 2. El grupo de respuesta de deseabilidad social es la tendencia a dar una calificación alta a un examinado en un rasgo sólo porque obtuvo una calificación alta en otro rasgo.

C. *Aparejamiento*

Instrucciones: Escriba la letra correspondiente al nombre correcto, de la lista que aparece en la segunda columna, en el espacio apropiado de la línea al margen de la primera columna.

_____ 1. prueba colectiva de inteligencia	A. Binet
_____ 2. prueba individual de inteligencia	B. Darwin
_____ 3. inventario de intereses	C. Galton
_____ 4. inventario de personalidad	D. Otis
_____ 5. correlación producto-momento	E. Pearson
_____ 6. pruebas sensoriomotrices	F. Rorschach
	G. Spearman
	H. Strong
	I. Woodworth

D. *Opción múltiple*

Instrucciones: Escriba la letra de la opción correcta en el espacio del margen al lado del reactivo.

_____ 1. Los adverbios como *nunca, en ocasiones* y *siempre*, que revelan la respuesta a un examinado que no tiene información sobre la materia del reactivo, se llaman
   a. generalidades brillantes
   b. adverbios de enlace
   c. grupos de respuesta
   d. determinantes específicos

_____ 2. Jimmy, quien tiene 8 años 4 meses de edad, obtiene una calificación de edad mental de 9 años 5 meses. ¿Cuál es su razón CI en la prueba?
   a. 88
   b. 90
   c. 113
   d. 120

simple adivinación. Sin embargo, las preguntas de ensayo pueden ser tan generales que se interpreten de manera muy diferente por distintas personas. Además, el número de preguntas de ensayo que pueden ser respondidas con respuestas de media página en una clase típica de 50 minutos (aproximadamente seis) puede ser insuficiente para determinar el conocimiento que tiene una persona de la materia de la prueba. No deberíamos esperar que las respuestas fueran tan inclusivas o detalladas como las requeridas por los reactivos del cuadro 2.1, pero podrían alcanzar cierta comprensión relativamente profunda del material. Otras desventajas de las pruebas de ensayo es que son susceptibles al engaño por parte de individuos con facilidad de palabra que no cuentan con información; además, su calificación es subjetiva y toma mucho tiempo.

Un profesor de historia informó haber aplicado una prueba de ensayo que incluía la pregunta: "¿Cuáles fueron las causas y las consecuencias de la Batalla de Hastings?" Y un estudiante apático, cuya preparación sobre la historia de Inglaterra no incluía los acontecimientos anteriores al siglo XIV, empezó a responder la pregunta con la afirmación: "No puedo comentar sobre la Batalla de Hastings, pero dirijamos nuestra atención hacia la Guerra de los Cien Años". Éste es un ejemplo bastante ostensible de la tendencia que los examinados no informados tienen a responder una pregunta ligeramente diferente de la que fue planteada para enfatizar lo que saben, en lugar de lo que no saben. Una manera de enfrentar este problema, aunque laborioso para los que presentan la prueba y para los que la califican, es el famoso procedimiento chino que consiste en hacer que los estudiantes escriban todo lo que saben del tema. Es posible que lo medido por esa prueba sea la susceptibilidad a la fatiga más que el conocimiento general.

Como regla, no deberán usarse reactivos de ensayo cuando el mismo conocimiento o habilidad puedan ser evaluados con reactivos objetivos. Si se plantean preguntas de ensayo, la persona que redacta los reactivos debe tratar de hacer las preguntas de manera tan objetiva como sea

CUADRO 2.1

¿UN EXAMEN FINAL GLOBAL PARA LOS ALUMNOS UNIVERSITARIOS DEL ÚLTIMO AÑO?

1. Describa la historia del pontificado desde sus orígenes hasta el presente, concentrándose especialmente, pero no de manera exclusiva, en su impacto social, político, económico, religioso y filosófico en Europa, Asia, África y América.
2. Con base en el conocimiento que tenga usted de sus obras, evalúe la estabilidad emocional, el grado de ajuste y las frustraciones reprimidas de Alejandro de Afrodisias, Ramsés II, Gregorio de Nicea y Hammurabi. Apoye sus respuestas con citas del trabajo de cada uno de estos hombres, mencionando las referencias apropiadas.
3. Desarrolle un plan realista para reducir la deuda nacional. Identifique los efectos de su plan sobre el cubismo, la controversia donatista y la teoría de las ondas de la luz. Bosqueje un método para prevenir esos efectos. Critique este método desde todos los puntos de vista posibles. Señale las deficiencias en su punto de vista, según lo demuestra su respuesta a la pregunta anterior.
4. Bosqueje el desarrollo del pensamiento humano. Estime su relevancia y compárelo con el desarrollo de cualquier otra clase de pensamiento —animal o alienígena.
5. Suponga que 2 500 aborígenes amotinados y enloquecidos están asaltando el aula. ¿Cómo los calmaría usted? Puede usar cualquier idioma antiguo excepto el latín o el griego y cualquier técnica no verbal diferente a la violencia.
6. Tome una posición a favor o en contra de la lógica y la verdad. ¿Cómo probaría la validez de su posición sin involucrar a nadie más o sin poner en peligro su propia salud?

posible. Esto puede lograrse al (1) definir la tarea y redactar los reactivos de manera clara, por ejemplo, pedir a los examinados que comparen y expliquen en lugar de que discutan; (2) usar un número pequeño de reactivos que deberán responder todos los examinados; (3) estructurar los reactivos de forma que los expertos en la materia estén de acuerdo en que puede demostrarse que una respuesta es mejor que otra, y (4) hacer que los examinados respondan a cada reactivo en una hoja por separado.

Reactivos de respuesta corta, de verdadero y falso y de aparejamiento

Los reactivos objetivos no se limitan a los cuatro tradicionales (respuesta corta o completamiento, verdadero y falso, aparejamiento y opción múltiple), pero éstos son los más populares. Entre las ventajas atribuidas a las pruebas objetivas está el que pueden calificarse de manera fácil e imparcial y que, como se requiere menos tiempo para responder a cada reactivo, puede hacerse un muestreo más amplio del contenido que en las pruebas de ensayo. Al preparar las pruebas objetivas debe tenerse cuidado de lograr que los reactivos resulten claros, precisos y gramaticalmente correctos. Deben escribirse en un lenguaje adecuado para el nivel de lectura de las personas a las que se dirigen. Debe incluirse en el reactivo toda la información y los requerimientos necesarios para seleccionar una respuesta razonable, omitiendo las palabras y frases no funcionales o estereotipadas.

Resulta tentador elaborar reactivos objetivos mediante la copia literal de afirmaciones de un texto o de otras fuentes, pero esta práctica sólo enfatiza la memoria. Las personas que redactan reactivos también deben tener cuidado de no incluir claves para las respuestas correctas y evitar los reactivos interrelacionados o entrelazados. Dos reactivos están *interrelacionados* cuando el planteamiento de uno proporciona una señal para la respuesta del otro. Dos reactivos están *entrelazados* cuando es necesario conocer la respuesta a uno de ellos para llegar a la respuesta correcta del otro.

Reactivos de respuesta corta. Un reactivo de respuesta corta o de completamiento plantea una tarea tipo fuente, en la cual se requiere que los examinados completen o llenen uno o más espacios en blanco de una afirmación incompleta con las palabras o frases correctas, o que den una respuesta breve a una pregunta. En términos de la longitud de la respuesta elaborada, los reactivos de respuesta corta caen entre los reactivos de ensayo y los de reconocimiento. Los reactivos de respuesta corta se encuentran entre los más sencillos de elaborar, y requieren que los examinados proporcionen la respuesta correcta en lugar de simplemente reconocerla. Aunque son especialmente útiles para evaluar el conocimiento de la terminología, los reactivos de respuesta corta tienen serias limitaciones: son inapropiados para medir objetivos instruccionales complejos y, debido a que puede haber más de una respuesta correcta, la calificación no siempre es por completo objetiva.[1]

Al elaborar reactivos de respuesta corta deberán seguirse las siguientes directrices:

1. Las preguntas directas son preferibles a las afirmaciones incompletas.
2. Plantee los reactivos de forma que las respuestas sean breves y no ambiguas.
3. Si se utiliza una afirmación incompleta, coloque el espacio en blanco al final de la afirmación.

[1]Un tipo de reactivo de completamiento diseñado para evaluar la habilidad de lectura es la *técnica cloze*. En este procedimiento se instruye a los individuos para que reemplacen las palabras faltantes que han sido borradas al azar en determinados párrafos. Una medida de la habilidad de lectura del individuo es el grado en que puede llenar correctamente los espacios en blanco y dar así sentido a los pasajes.

4. Haga que todos los espacios en blanco sean de la misma extensión.
5. Evite usar múltiples espacios en blanco en el mismo reactivo, en especial si tornan poco claro el significado de la tarea.
6. Indique las unidades en que deben expresarse las respuestas numéricas.

Reactivos de verdadero y falso. Uno de los tipos de reactivos que es más sencillo de elaborar, pero probablemente el más criticado por los examinadores profesionales, es el de verdadero y falso. Los reactivos de verdadero y falso pueden escribirse y leerse con rapidez y, por ende, permiten un muestreo amplio del contenido de la materia. Un defecto notorio de los reactivos de verdadero y falso es que a menudo se interesan en información trivial o se elaboran copiando afirmaciones literales de un texto. En consecuencia, se dice que alientan la memorización y así encaminan mal los esfuerzos por aprender. Otra crítica a estos reactivos es que a menudo son ambiguos y no pueden usarse para medir objetivos instruccionales más complejos. Además, debido a que la calificación total en una prueba de este tipo puede ser afectada por la tendencia del examinado a adivinar cuando tiene dudas o a estar de acuerdo (o en desacuerdo), la precisión de la calificación puede ser cuestionable.[2]

En promedio, los examinados obtendrán un 50% de aciertos en los reactivos de verdadero y falso simplemente por adivinar. Las calificaciones pueden ser infladas todavía más cuando los reactivos contienen *determinantes específicos* —palabras como *todos, siempre, nunca* y *sólo*—, los cuales indican que la afirmación probablemente es falsa, o palabras como *a menudo, en ocasiones* y *usualmente*, sugerentes de que la afirmación es verdadera.

A pesar de esos defectos, los reactivos de verdadero y falso no tienen que ser triviales o ambiguos o encaminar mal el aprendizaje. En defensa de los reactivos de verdadero y falso, Ebel (1979) afirma que el grado de dominio que tienen los estudiantes en un área particular del conocimiento es indicado por su éxito al juzgar la veracidad o falsedad de proposiciones relacionadas con él (p. 112). Él ha considerado que tales proposiciones son expresiones del conocimiento verbal, que es la esencia del logro educativo.

La defensa que hizo Ebel de los reactivos de verdadero y falso puede ser cuestionada, pero no se cuestiona el hecho de que estos reactivos, bien diseñados, pueden medir más que la simple memoria. Por ejemplo, al incluir dos conceptos, condiciones o eventos en un reactivo de verdadero y falso, el examinador puede preguntar si es verdad que tienen una relación de moderada a fuerte (Diekhoff, 1984). Otras posibilidades son preguntar si (1) un concepto, condición o evento implica o es una consecuencia de otro evento; (2) un concepto, condición o evento es un subconjunto, ejemplo o categoría de otro evento, (3) ambos conceptos, condiciones o eventos son verdaderos. Dichos reactivos pueden medir la comprensión así como el conocimiento significativo de conceptos y eventos.

Cualesquiera que sean los objetivos de una prueba de verdadero y falso, al elaborar reactivos de este tipo es recomendable atender las siguientes sugerencias:

1. Asegúrese de que las afirmaciones planteen asuntos importantes (no triviales).
2. Establezca afirmaciones relativamente cortas, y verdaderas o falsas sin lugar a dudas.
3. Evite los reactivos planteados de manera negativa, especialmente los que contienen doble negación.
4. Evite los reactivos ambiguos y capciosos.

[2]La tendencia a estar de acuerdo cuando se tiene duda (o conformidad) es un grupo de respuesta. Los *grupos de respuestas* son las tendencias por parte de los examinados a responder a los reactivos de una prueba de acuerdo con su forma, es decir, a la manera en que están planteados, en lugar de hacerlo con base en su contenido.

5. Como regla, evite los determinantes específicos. Si se usan determinantes específicos para hacer que se equivoquen las personas sin conocimientos, pero hábiles para presentar pruebas, deben incluirse en las afirmaciones verdaderas tan a menudo como en las falsas.
6. En las afirmaciones de opinión, cite la fuente.
7. Haga que las afirmaciones verdaderas y las falsas sean aproximadamente de la misma longitud, y que el número de afirmaciones verdaderas sea aproximadamente igual al de las falsas. Puede argumentarse que, dado que los reactivos falsos tienden a discriminar más que los reactivos verdaderos, el número de afirmaciones falsas debería ser mayor que el de afirmaciones verdaderas. Sin embargo, si el maestro sigue esta práctica en pruebas sucesivas, los estudiantes pueden darse cuenta de ello y comenzar a responder "falso" cuando tengan duda acerca de la respuesta.
8. Asegúrese de que las respuestas erróneas sean más atractivas planteando los reactivos de tal manera que la lógica superficial, los errores populares o los determinantes específicos sugieran que las respuestas erróneas son correctas. Las afirmaciones falsas que parecen verdaderas también pueden hacer que se equivoquen los examinados sin conocimientos.

Reactivos de aparejamiento. Tanto los reactivos de verdadero y falso como los de opción múltiple son, en cierto sentido, variedades de los reactivos de aparejamiento. En estos tres tipos de reactivos, un conjunto de opciones de respuesta se equipara con un conjunto de opciones de estímulo (premisas). La distinción es que los reactivos de verdadero y falso y los de opción múltiple tienen sólo una premisa (el *tronco* del reactivo) y dos o más opciones de respuesta, mientras que los reactivos de aparejamiento tienen múltiples premisas y múltiples opciones de respuesta. La tarea del examinado en un reactivo de aparejamiento es acoplar las opciones de respuesta con la premisa correcta. El aparejamiento usualmente es de uno a uno (una respuesta por premisa), pero también puede ser de una respuesta a varias premisas, de varias respuestas a una premisa, o de varias respuestas a varias premisas. Por supuesto, debe informarse a los examinados cuál de esos procedimientos se aplica en un reactivo particular.

Los reactivos de aparejamiento son más sencillos de elaborar y cubren el material de manera más eficiente que muchos otros tipos de reactivos; por desgracia, usualmente sólo miden la memorización de acontecimientos.[3] Además, la necesidad de hacer que las opciones sean homogéneas (que todas las opciones de respuesta sean del mismo tipo, como fechas, lugares o nombres) limita el tipo de material que puede adaptarse a un marco de aparejamiento. A continuación se presentan algunos lineamientos para elaborar reactivos de aparejamiento:

1. Ordene la premisa y las opciones de respuesta en un formato claro y lógico de columnas, con las premisas en la columna izquierda y las opciones de respuesta en la columna derecha.
2. Use entre seis y quince premisas, con dos o tres opciones de respuesta más que premisas.
3. Numere las premisas de manera sucesiva, y coloque letras (a, b, c, etc.) antes de las opciones de respuesta.
4. Especifique con claridad las bases para realizar el aparejamiento.
5. Coloque todo el reactivo en una sola página.

Un tipo especial de reactivo de aparejamiento es el *reactivo de reordenamiento*, en el cual se requiere que los examinados clasifiquen un número fijo de categorías predeterminadas. En un

[3]Al menos un estudio encontró que los reactivos de aparejamiento pueden diseñarse para ser iguales o incluso superiores a los de opción múltiple como medidas tanto del dominio del contenido de interés como de las actitudes de las personas que presentan la prueba (Shaha, 1984).

tipo particular de reordenamiento conocido como *reactivo de rango*, los individuos reordenan un conjunto de opciones en orden de la primera a la última (o de la más alta a la más baja).

Reactivos de opción múltiple

No se sabe quién elaboró el primer reactivo de opción múltiple para una prueba, pero desde el punto de vista de la evaluación psicológica fue algo fortuito.[4] Los reactivos de opción múltiple son los más versátiles de todos los reactivos objetivos, ya que pueden usarse para medir logros de aprendizaje simples y complejos en todos los niveles y en todas las áreas temáticas. Aunque los reactivos de respuesta de ensayo demandan mayor habilidad de organización que la selección de respuestas a los reactivos de opción múltiple, responder de manera correcta a un reactivo de opción múltiple bien preparado requiere buena habilidad para discriminar y no sólo capacidad para reconocer o recordar la respuesta correcta. Las calificaciones en los reactivos de opción múltiple también son menos afectadas por la adivinación y por otros grupos de respuesta que las calificaciones en otros reactivos objetivos. Además, puede obtenerse información de diagnóstico útil a partir de un análisis de las opciones incorrectas (*distractores*) seleccionadas por los examinados.

Entre los defectos de los reactivos de opción múltiple están que (1) los buenos son difíciles de elaborar, en especial aquellos en los que todas las opciones resulten igualmente atractivas para los examinados que no conocen la respuesta correcta; (2) enfatizan el reconocimiento más que el recuerdo y la organización de la información, y (3) requieren más tiempo para la respuesta y pueden muestrear el dominio temático de manera menos adecuada que los reactivos de verdadero y falso. También se ha argumentado, pero no demostrado, que las pruebas de opción múltiple favorecen a los lectores sagaces, hábiles y rápidos, y penalizan a los personas más reflexivas y que piensan con más profundidad (Hoffman, 1962).

En el cuadro 2.2 se presentan lineamientos para facilitar la elaboración de reactivos de opción múltiple de alta calidad. Tales lineamientos son sobre todo producto de la lógica y de la experiencia, más que de la investigación, y su seguimiento no garantiza la elaboración de buenas pruebas de opción múltiple. La elaboración de buenos reactivos depende mucho o más que del conocimiento de la materia de la prueba, de la comprensión de lo que los estudiantes deberían saber y de lo que es poco probable que sepan acerca de la materia, y del arte o habilidad de plantear preguntas. Incluso cuando los lineamientos no se siguen con precisión, los reactivos de opción múltiple tienden a ser bastante sólidos en su capacidad para medir el conocimiento y la comprensión.

Elaboración de distractores. Un factor crucial en la determinación de la efectividad de los reactivos de opción múltiple es la selección o elaboración de los elementos distractores (las opciones incorrectas). Para la selección de reactivos puede emplearse una aproximación racional o una empírica. El enfoque *racional* demanda a la persona que elabora la prueba formular juicios personales concernientes a qué distractores son apropiados. En contraste, el enfoque *empírico* consiste en seleccionar distractores de entre las respuestas incorrectas más populares a los troncos de los reactivos aplicados en afirmaciones abiertas-cerradas. No hay consenso acerca de qué método da lugar a los mejores distractores, pero el juicio del examinador parece ser al menos tan efectivo como la aproximación empírica (Hanna y Johnson, 1978; Owens, Hanna y Coppedge, 1970).

[4]Se acredita a Arthur Otis haber sido pionero en el uso del formato de reactivo de opción múltiple en las pruebas colectivas de inteligencia. Los primeros instrumentos publicados que emplearon este formato fueron las Pruebas autoaplicables de Otis de habilidad mental (1916-1917).

CUADRO 2.2
LINEAMIENTOS PARA ELABORAR REACTIVOS DE OPCIÓN MÚLTIPLE

1. Debe utilizarse como tronco una pregunta o una afirmación incompleta, pero se prefiere el formato de pregunta. Si el tronco es una afirmación incompleta, coloque el espacio en blanco al final de la afirmación.
2. Establezca claramente el problema específico de la pregunta o afirmación incompleta en el tronco y a un nivel de lectura apropiado para los examinados, pero evite tomar preguntas o afirmaciones literales de los textos.
3. Coloque la mayor parte del reactivo en el tronco. Es ineficiente repetir las mismas palabras en cada opción y a los examinados les resulta menos difícil revisar las opciones más cortas.
4. Emplee preguntas de opinión con moderación; cuando las utilice, cite la fuente de la opinión.
5. Cuatro o cinco opciones son típicas, pero también pueden escribirse buenos reactivos que tengan sólo dos o tres opciones. Con los estudiantes de los primeros grados, tres opciones son preferibles a cuatro o cinco. Haladyna y Downing (1993) concluyeron que tres opciones pueden ser adecuadas para la mayor parte de las pruebas de habilidad y rendimiento.
6. Si las opciones tienen un orden natural, como fechas o edades, es aconsejable disponerlas en ese orden. De otro modo, ordénelas aleatoria o alfabéticamente (siempre que la alfabetización no proporcione señales para la respuesta correcta).
7. Haga que todas las opciones sean aproximadamente de la misma extensión, que sean gramaticalmente correctas y apropiadas en relación con el tronco. Sin embargo, no deje que el tronco revele la opción correcta por medio de asociaciones verbales u otras señales.
8. Haga que todas las opciones sean plausibles para los examinados que no conocen la respuesta correcta, pero haga que sólo una opción sea la correcta o "la mejor". Los errores populares o las afirmaciones que sólo son parcialmente correctas son buenos distractores.
9. Al elaborar cada distractor, plantee una razón por la cual los examinados que no conocen la respuesta correcta podrían seleccionarlo.
10. Evite, o al menos minimice, el uso de expresiones negativas como "no" en el tronco o las opciones.
11. Aunque cierta cantidad de novedad e incluso de humor es apropiada y puede servir para interesar y motivar a los examinados, no deben usarse reactivos y opciones ambiguos o capciosos.
12. Use con moderación las expresiones: ninguno de los anteriores, todos los anteriores, o más de uno de los anteriores. Además, evite el uso de determinantes específicos como: siempre o nunca.
13. Coloque las opciones en un formato apilado (párrafo) en lugar de hacerlo en tándem (una tras otra); use números para designar los reactivos y letras para las opciones.
14. Prepare el número correcto de reactivos para el grado o nivel de edad que se pondrá a prueba, haciendo que cada reactivo sea independiente de otros reactivos (que no se entrelacen o se interrelacionen).
15. Haga que los niveles de dificultad sean tales que el porcentaje de examinados que responden a un reactivo de manera correcta esté aproximadamente a la mitad entre el porcentaje de azar (adivinación aleatoria) y el 100 por ciento: % correcto = 50(k + 1)/k, donde k es el número de distractores por reactivo.

Elaboración de reactivos complejos. Los diseñadores de pruebas por lo general tienen más dificultad para elaborar reactivos que midan la comprensión y el pensamiento que los que miden el conocimiento directo de la materia. Se han propuesto varias formas de redactar reactivos objetivos que evalúen objetivos instruccionales más complejos. Opciones como: todas las anteriores, ninguna de las anteriores, dos de las anteriores y todas salvo una de las anteriores, pueden tornar más difícil la elección de un examinado. Dicha elección también puede complicarse haciendo que todas las opciones sean correctas (o incorrectas) y pidiendo a los examinados que seleccionen la mejor o la más apropiada para cada reactivo. Otras maneras de hacer más difícil la decisión de un

examinado son: (1) incluir reactivos de respuesta múltiple en los cuales números variables de opciones sean correctos y el examinado deba indicar qué opciones (si las hay) son correctas o incorrectas; (2) hacer que los examinados seleccionen una respuesta y la mejoren o escriban una breve justificación de la misma, y (3) pedir a los examinados que identifiquen el planteamiento correcto (como una ecuación o método de solución) en tareas de resolución de problemas.

En el cuadro 2.3 se ilustran otros procedimientos para incrementar la complejidad de los reactivos de opción múltiple. Todas esas técnicas están diseñadas para hacer que la selección de la opción correcta sea un proceso reflexivo y analítico, en el cual se pongan en práctica varias capacidades cognoscitivas en lugar de sólo la memoria. Por último, el uso de un formato de conjunto de problemas, en el cual dos o más reactivos de opción múltiple se relacionan con la misma ilustración, gráfica, pasaje o escenario, se ha vuelto popular en los exámenes de acreditación o certificación (Hambleton, 1996).

Uso de computadoras en la elaboración de pruebas

Las aplicaciones más comunes de las computadoras en la elaboración de pruebas consisten en programas de procesamiento de textos para ayudar en la mecanografía de los reactivos, la formación, la revisión de errores de ortografía y de sintaxis, etc. La elaboración de pruebas es facilitada aún más por una combinación del procesador de textos y programas de gráficos que apoyan la preparación de pruebas compuestas por palabras e ilustraciones. Esos programas contienen bancos de reactivos a los cuales se puede tener acceso ingresando ciertas palabras clave que indican el contenido y las características psicométricas deseadas en la prueba. Los bancos de reactivos, de los que pueden seleccionarse y recuperarse los reactivos al diseñar las pruebas, están disponibles con los editores de libros de texto como complementos para determinadas obras.

Los redactores de reactivos de prueba basados en la computadora, algoritmos de especificación de dominio para generar reactivos de prueba, y enfoques basados en la lingüística o el aprendizaje de conceptos para la redacción de reactivos, pueden proporcionar procedimientos más eficientes y precisos para la elaboración de reactivos de prueba (Herman, 1994). En la actualidad la preparación de buenos reactivos de prueba es tanto un arte como una ciencia.

FORMACIÓN Y REPRODUCCIÓN DE UN TEST

Una vez que se han preparado los reactivos para una prueba, es aconsejable hacer que los revisen y editen personas conocedoras. Incluso los esfuerzos más concienzudos no necesariamente producen una buena prueba, y un amigo o asociado con frecuencia puede detectar errores y hacer sugerencias valiosas para mejorar los reactivos.

Suponiendo que el diseñador de la prueba ha elaborado un número suficiente de reactivos satisfactorios, antes de formar una prueba deben tomarse decisiones finales concernientes a varios asuntos:

1. ¿Es la longitud de la prueba es apropiada para los límites de tiempo?
2. ¿Cómo deberán agruparse u ordenarse los reactivos en las páginas del cuadernillo de la prueba?
3. ¿Deben marcarse las respuestas en el cuadernillo de la prueba o se utilizará una hoja especial de respuestas?
4. ¿Cómo se reproducirán el cuadernillo de la prueba y la hoja de respuestas?
5. ¿Qué información debe incluirse en las instrucciones de la prueba?

CUADRO 2.3

ALGUNAS FORMAS COMPLEJAS DE REACTIVOS DE OPCIÓN MÚLTIPLE

1. *Clasificación*. El examinado clasifica a una persona, objeto o condición en una de varias categorías diseñadas en el tronco.

   Jean Piaget se caracteriza mejor como un psicólogo _____ .

   a. clínico
   b. del desarrollo
   c. psicómetra
   d. social

2. *Condiciones si-entonces*. El examinado debe determinar la consecuencia correcta de una o más condiciones presentes.

   Si la varianza verdadera de una prueba se incrementa, pero la varianza de error permanece constante, ¿cuál de las siguientes situaciones ocurrirá?

   a. la confiabilidad aumentará
   b. la confiabilidad disminuirá
   c. la varianza observada disminuirá
   d. ni la confiabilidad ni la varianza observada cambiarán

3. *Condiciones múltiples*. El examinado utiliza las condiciones o afirmaciones presentadas en el tronco para derivar una conclusión.

   Si la media de una prueba es 59 y su desviación estándar es 2, ¿cuál es la calificación *z* de María si su calificación cruda en la prueba es 60?

   a. –2.00
   b. –.50
   c. .50
   d. 2.00

4. *Verdadero y falso múltiple*. El examinado decide si una, todas o ninguna de las dos o más condiciones o afirmaciones presentadas en el tronco es(son) correcta(s).

   ¿Es cierto que (1) Alfred Binet fue el padre de las pruebas de inteligencia, y (2) su primera prueba de inteligencia se publicó en 1916?

   a. ambas 1 y 2
   b. 1 pero no 2
   c. 1 no pero 2 sí
   d. ni 1 ni 2

5. *Falta de correspondencia*. El examinado indica cuál opción no pertenece al mismo grupo que las otras.

   ¿Cuál de los siguientes nombres no corresponde con los otros?

   a. Alfred Adler
   b. Sigmund Freud
   c. Carl Jung
   d. Carl Rogers

6. *Relaciones y correlatos*. El examinado determina la relación entre dos conceptos e indica cuál de ellos (a, b, c, d, etc.) se relaciona con un tercer concepto de la misma manera que los dos primeros conceptos se relacionan entre sí.

   La media es a la desviación estándar como la mediana es a:

   a. la desviación promedio
   b. el rango inclusivo
   c. el rango semiintercuartilar
   d. la varianza

Extensión de la prueba

La decisión de cuántos reactivos incluir en una prueba depende de los límites de tiempo, del grado y nivel de lectura de los examinados, y de la extensión y dificultad de los reactivos. Los reactivos cortos y/o los que sólo requieren memorización de acontecimientos pueden responderse en menos tiempo que los más largos, donde son necesarios cálculos laboriosos y/o razonamiento abstracto. La experiencia previa con reactivos del mismo tipo general que los incluidos en una prueba ayudará a determinar si los límites de tiempo son apropiados. En las pruebas de dificultad moderada aplicadas a partir del nivel de las escuelas secundarias, una buena regla empírica es conceder un minuto por cada reactivo de opción múltiple o de respuesta corta y un minuto por cada dos reactivos de verdadero y falso. De este modo, una prueba de 50 reactivos de opción múltiple o de respuesta corta y una de 100 reactivos de verdadero y falso suelen ser apropiadas para un periodo de clase típico de 50 minutos en el nivel de secundaria. Cinco o seis preguntas de ensayo que requieren respuestas de media página pueden ser respondidas en este mismo periodo. A menos que los reactivos sean muy largos o sumamente difíciles, al menos 90% de los estudiantes en un grupo típico de secundaria podrán terminar la prueba en el tiempo asignado. La extensión de la prueba y los límites de tiempo necesitarán ajustarse hacia abajo o hacia arriba cuando se examine a alumnos de escuela primaria o a estudiantes de universidad.

Existen, por supuesto, diferencias entre los estudiantes en cuanto al tiempo que requieren para terminar una prueba. Puede esperarse que aquellos con más conocimientos o habilidades en la materia de la prueba terminen primero, pero no siempre sucede así. Los estudiantes menos informados pueden simplemente adivinar o "rendirse" y entregar la prueba antes del tiempo límite cuando se permita hacerlo. Además, los hábitos de presentación de pruebas de los examinados con altas calificaciones pueden llevarlos a revisar los reactivos de la prueba en varias ocasiones para estar seguros de que no pasaron algo por alto o lo interpretaron mal. Ciertos estudiantes, con altas y bajas calificaciones, también habrán escuchado que es más probable que sus respuestas iniciales sean las correctas, y por lo tanto no es buena idea perder tiempo reconsiderando la primera elección. Todos esos factores hacen difícil predecir cuánto tiempo le tomará a un alumno determinado terminar una prueba. Todo depende de una interacción compleja entre la preparación, la personalidad y el estado emocional y físico del estudiante, de la naturaleza y dificultad del material de la prueba, y del ambiente del examen (ruido y otras distracciones, conducta del examinador o supervisor, etcétera). Es probable que quien administre la prueba pueda hacer que el tiempo real dedicado a resolverla sea más uniforme al pedir que los examinados permanezcan en sus asientos después de terminarla, pero aún así puede haber diferencias sustanciales en el tiempo que necesitan los examinados para completar la prueba.

Ordenamiento de los reactivos

En lo que respecta al ordenamiento de las opciones en los reactivos de opción múltiple, se ha dicho que los examinados muestran preferencias por la posición de las opciones, y cuando no están seguros de la respuesta es más probable que elijan ciertas opciones (digamos *b* y *c*) que otras (*a* y *d*). Aunque la investigación no ha logrado demostrar que estas preferencias tengan un efecto significativo en las calificaciones de una prueba (Jessell y Sullins, 1975; Wilbur, 1970), es aconsejable ordenar los reactivos de opción múltiple y los de verdadero y falso de forma que las respuestas no sigan un patrón. Ordenar las opciones para los reactivos de opción múltiple en orden alfabético puede ser satisfactorio, pero una mejor estrategia es aleatorizar el orden de las opciones dentro de los reactivos. Esto asegurará que al menos la persona que elabora la prueba no tenga ninguna inclinación al ordenar las opciones correctas. Por supuesto, cuando se usan, op-

ciones como: todas las anteriores y ninguna de las anteriores, éstas deben colocarse en la última posición.

En los reactivos de aparejamiento o reordenamiento, a los examinados les resulta más conveniente y la calificación se facilita si todas las premisas y opciones de respuesta se colocan en la misma página. Colocar los reactivos de respuesta corta en grupos de cinco o algo así también puede reducir los errores al presentar y calificar una prueba. Por último, debe proporcionarse espacio suficiente para responder los reactivos de respuesta corta y los de ensayo, sea que las respuestas se escriban en el cuadernillo de la prueba o en una hoja por separado.

Con relación al esquema de la prueba como un todo, puede esperarse que la tarea de los examinados se haga más sencilla si se agrupan juntos los reactivos del mismo tipo (opción múltiple, verdadero y falso, etc.) y los que tratan del mismo tema. Es cierto que ordenar los reactivos en grupos de acuerdo con el tipo o tema puede simplificar la preparación, aplicación y calificación de la prueba, pero no hay evidencia de que esta práctica mejore las calificaciones del instrumento. En las pruebas que contienen reactivos objetivos y de ensayo, estos últimos suelen colocarse al final, ya que suelen requerir más tiempo y diferentes procesos de pensamiento que los primeros.

Otra suposición razonable es que las calificaciones de la prueba serán más altas si se ordenan subconjuntos de reactivos del más fácil al más difícil. Se supone que el éxito al responder los reactivos más sencillos crea expectativas favorables de éxito, y que ello anima a los examinados a poner más empeño en los reactivos más difíciles. Sin embargo, los hallazgos de la investigación no siempre han confirmado esta suposición (Allison, 1984; Gerow, 1980; Klimko, 1984). Un reactivo sencillo ocasional puede mejorar el desempeño en los reactivos subsecuentes, pero, en general, ordenar los reactivos en orden de dificultad parece tener poco efecto sobre las calificaciones globales. Las excepciones a esta conclusión son las pruebas de velocidad (Plake, Ansorge, Parker y Lowry, 1982) o las muy difíciles (Green, 1984; Savitz, 1985). En una prueba de velocidad o en una que es muy difícil, colocar los reactivos más difíciles al final de la prueba parece mejorar un tanto las calificaciones.

Una conclusión lógica de los hallazgos de la investigación sobre los efectos del ordenamiento en los reactivos de acuerdo con el nivel de dificultad es que, al elaborar pruebas que no son de velocidad desde fáciles hasta de dificultad moderada, los diseñadores harían bien en preocuparse menos por el ordenamiento de los reactivos e interesarse más en asegurarse de que estén bien escritos y midan lo que se supone deben medir. Cuando una prueba es muy difícil o de velocidad, colocar los reactivos en orden de los más fáciles a los más difíciles puede asegurar el uso más eficiente del tiempo del examinado, así como mejorar la motivación y, por consiguiente, dar por resultado calificaciones más altas.

Hojas de respuestas

Para la mayoría de las pruebas que se administran en un aula, en especial en los primeros grados, es aconsejable hacer que los estudiantes marquen o escriban sus respuestas en el cuadernillo de la prueba (Airasian y Terrasi, 1994). Esto genera menos errores al indicar las respuestas. En los reactivos objetivos, también facilita la calificación si se requiere que los examinados escriban las letras o respuestas apropiadas en los espacios marginales situados a la izquierda de las preguntas.

Las hojas de respuestas por separado, que son más fáciles de calificar, pueden usarse a partir de los últimos años de la escuela elemental. Si la prueba se va a calificar con una máquina deberán usarse hojas de respuestas distribuidas comercialmente. En dichas hojas, los examinados responden colocando en un círculo o en un espacio al lado del número del reactivo el número o letra correspondiente. Si la prueba va a calificarse de manera manual, el profesor puede

preparar fácilmente una hoja de respuestas y duplicarla. Una hoja de respuestas para una prueba de 75 reactivos de opción múltiple puede tener el siguiente formato:

1. a b c d e	**26.** a b c d e	**51.** a b c d e
2. a b c d e	**27.** a b c d e	**52.** a b c d e
. . .	. . .	. . .
25. a b c d e	**50.** a b c d e	**75.** a b c d e

Se indica a los examinados que marquen la letra correspondiente a la respuesta correcta para cada reactivo. También se dispone de hojas de respuestas SCANTRON que pueden ser calificadas por una máquina o a mano.

Toda institución educativa tiene recursos que facilitan la reproducción de materiales escritos o impresos para su uso en el aula. Las máquinas fotocopiadoras pueden utilizarse para reproducir los cuadernillos de prueba en un formato de impresión por uno o ambos lados, en ocasiones a color. Si se va a usar el mismo tipo de hoja de respuestas para diferentes pruebas, puede imprimirse una gran cantidad en una sola operación de la máquina y almacenarse para otras aplicaciones de pruebas.

Instrucciones en los tests

Las instrucciones generales para una prueba de ensayo u objetiva que se aplica de manera simultánea a un grupo de personas se colocan al frente de la prueba, y las instrucciones específicas para cada parte de una prueba múltiple se colocan antes de la parte respectiva. Por lo general, resulta sensato mecanografiar las instrucciones en negritas de forma que sea menos probable que los examinados las salten o las pasen por alto. Como su planteamiento puede llegar a tener cierto efecto sobre las calificaciones obtenidas, las instrucciones deben ser precisas más que generales (Joncas y Standig, 1998). También es aconsejable que el examinador lea en voz alta las instrucciones globales si son inusuales o poco familiares para los examinados. En una prueba individual en la cual el examinador presenta cada tarea e interactúa de manera continua con el examinado, las instrucciones se dan en forma oral. Sea que se den de manera oral, impresa o en ambas formas, las instrucciones deben informar a los examinados acerca del propósito de la prueba (o reactivo), cómo deben indicarse las respuestas,[5] qué tipo de ayuda pueden esperar si no entienden algo, cuánto tiempo tienen para terminar la prueba, cómo se calificarán las respuestas, si es recomendable adivinar cuando se tenga duda, y cómo corregir una respuesta si cometieron un error. Las siguientes instrucciones generales para una prueba de rendimiento aplicada a un grupo son representativas:

> Escriba su nombre en la esquina superior derecha de la hoja de respuestas, pero no escriba en el cuadernillo de la prueba. Esta prueba está diseñada para evaluar su conocimiento y comprensión de estadística elemental. Son 50 reactivos y usted tendrá exactamente 50 minutos para completar la prueba. Indique su respuesta a cada reactivo llenando el espacio apropiado en la hoja de respuestas debajo de la letra que corresponde a la respuesta correcta. Su calificación en la prueba será igual al número de reactivos que haya respondido correctamente. Aunque la adivinación al azar no aumentará su calificación, si puede eliminar al menos una opción en un reactivo, es sensato hacer una con-

[5]Como el método de respuesta en las pruebas aplicadas por computadora puede no ser familiar para algunos examinados, debe asignarse tiempo suficiente para dar las instrucciones y mostrar cómo funciona el equipo. Además, los examinados deben ser supervisados durante la prueba para asegurarse de que están usando el equipo de manera apropiada.

jetura informada a partir de las opciones restantes. Debe tener tiempo suficiente para responder todos los reactivos y revisar sus respuestas. Si termina antes de tiempo, por favor permanezca sentado en silencio hasta que todos hayan terminado.

Cuando las instrucciones de una prueba se den de manera oral, deben leerse de forma lenta, clara y exactamente como aparecen impresas. Después de haber leído las instrucciones, debe permitirse a los examinados hacer preguntas, e independientemente de su trivialidad o redundancia aparente, deben responderse de manera paciente e informativa.

En las pruebas múltiples que constan de cierta variedad de temas y/o tipos de reactivos, puede ser necesario dar instrucciones específicas para cada parte. Las instrucciones que atañen a muchos de los mismos asuntos (cómo marcar las respuestas, cómo corregir los errores, si se pueden omitir respuestas o adivinar cuando se tenga duda) pueden variar con el tipo de reactivos objetivos. Las instrucciones para responder los reactivos de ensayo pueden incluir sugerencias acerca de cómo estructurar las respuestas (bosquejo, formato y cosas similares); cómo deben ser las respuestas largas; qué tanto peso de calificación se dará al contenido, forma, gramática, caligrafía y otros rasgos de las respuestas, y si debe intentarse responder a todas las preguntas, a un número selecto de éstas, o si algunas son obligatorias y otras opcionales.

PRUEBAS ORALES

Las pruebas orales se definen como una situación de evaluación en la cual los examinados responden de manera oral a las preguntas planteadas. Las preguntas pueden presentarse de manera oral, por escrito o de ambas formas. Las pruebas orales de rendimiento son más comunes en las instituciones educativas europeas que en Estados Unidos, donde la práctica de las pruebas orales declinó durante el siglo XX y es menos común en los grados superiores que en los inferiores.

A muchos estudiantes no les gustan las pruebas orales y sienten que son medidas injustas del conocimiento y la comprensión. Sin embargo, los maestros de expresión oral, arte dramático, inglés e idiomas extranjeros, a menudo deploran la falta de atención a la evaluación de las habilidades del lenguaje hablado y sienten que la consecuencia de semejante descuido es una ciudadanía que no puede hablar de manera correcta, comprensible y cómoda. Aunque muchos maestros de idiomas y de otras materias en las cuales es importante el desarrollo de las habilidades del habla admiten lo deseable de los ejercicios y evaluaciones orales, también se dan cuenta de que las pruebas orales no sólo son muy subjetivas sino que a menudo resultan ineficientes (Crowl y McGinitie, 1974; Platt, 1961).

Ventajas de las pruebas orales

Desde los primeros años del siglo XX, las pruebas orales de rendimiento se han venido percibiendo como carentes de eficiencia y rigor psicométrico. También se les ha criticado por requerir demasiado tiempo, proporcionar una muestra limitada de respuestas y por estar mal planeadas en la mayoría de los casos.

Sin embargo, a pesar de sus limitaciones, incluso los críticos de las pruebas orales admiten que éstas poseen algunas ventajas sobre las pruebas escritas. Una ventaja es la situación social interactiva que proporcionan, lo que permite evaluar cualidades personales como apariencia, estilo y manera de hablar. La situación cara a cara también hace poco probable la copia y quizá los engaños. Otras ventajas de las pruebas orales es que con frecuencia requieren respuestas a un

nivel intelectual más alto que las escritas, y proporcionan práctica en comunicación oral e interacción social. También alientan una revisión más cuidadosa del material de prueba y pueden ser terminadas en menos tiempo que exámenes escritos comparables. Los individuos que aplican pruebas orales pueden seguir los procesos de pensamiento de los examinados y localizar con más facilidad los límites de su conocimiento y comprensión de la materia. Esos límites pueden ser determinados pidiendo a los examinados que expliquen, defiendan o se esmeren en sus respuestas. Por último, el tiempo que se necesita para preparar y evaluar las respuestas orales puede ser menor que para una prueba escrita comparable (Glovrozov, 1974; Platt, 1961).

Las pruebas orales son especialmente apropiadas para los alumnos de primaria y para otros que experimentan dificultades en la lectura o escritura. Incluso en los niveles superiores puede estar justificada la aplicación de una prueba oral ocasional cuando el tiempo y/o los recursos para reproducir los materiales de prueba son escasos (Green, 1975). Los exámenes orales son cruciales en materias como expresión oral, idiomas y arte dramático.

Las entrevistas estructuradas que constan de preguntas y respuestas orales a menudo se realizan con solicitantes de puestos en organizaciones gubernamentales e industriales. Es frecuente que tales entrevistas se efectúen por teléfono cuando los solicitantes no pueden viajar al sitio del examen. En exámenes de este tipo es posible introducir cierta cantidad de estandarización y control planteando a todos los examinados las mismas preguntas, limitando el tiempo del que disponen para responder y registrando electrónicamente sus respuestas para reproducirlas y evaluarlas más tarde.

Pruebas orales contra pruebas escritas

El hecho de que las calificaciones en las pruebas orales de rendimiento sólo tengan correlaciones moderadas con las calificaciones en pruebas escritas comparables, sugiere que miden aspectos diferentes del rendimiento. En general, el conocimiento de hechos específicos puede ser determinado con mayor rapidez por las pruebas objetivas escritas, por lo que los exámenes orales no deben contener grandes cantidades de esos tipos de preguntas. Como sucede con las pruebas de ensayo, las pruebas orales son más apropiadas cuando las preguntas requieren de respuestas extensas.

Dado que los logros o las conductas evaluadas mediante pruebas orales son tan importantes como las mediciones de pruebas escritas, debería prestarse más atención a la principal fuente de error en las pruebas orales: los examinadores o evaluadores. Las personas que aplican pruebas orales deben poseer un conocimiento profundo de la materia y una conciencia muy aguda de las respuestas apropiadas. Además, las categorías usadas por los examinadores al describir o calificar las respuestas de los examinados deberían citar conductas observables específicas en lugar de conceptos vagos como *potencial creativo, carácter, habilidad general* o *efectividad interpersonal*. Estos conceptos indefinidos, y que quizá no puedan definirse, no son medidos más fácilmente por las pruebas orales que por las escritas.

PRUEBAS DE DESEMPEÑO

Las pruebas de lápiz y papel son las más eficientes y objetivas de todos los tipos de pruebas, pero por lo regular sólo proporcionan información indirecta acerca de la habilidad de una persona para hacer o fabricar algo. El conocimiento de la materia puede demostrarse de manera bastante minuciosa en un periodo corto por medio de una prueba de ensayo, una de opción múltiple u otra prueba escrita. Sin embargo, poseer un bagaje de información acerca de un tema o ser capaz de explicar cómo hacer algo no es lo mismo que usar la información o destreza en situacio-

nes prácticas. En alguna ocasión el autor condujo un taller de relaciones humanas con un grupo de supervisores de una línea de ensamblaje. Aunque todos los supervisores salieron bien en las pruebas escritas del material presentado en el taller y coincidieron en que un enfoque democrático hacia la supervisión era superior a uno autoritario, la mayoría reanudó su conducta autoritaria en la supervisión al regresar a la línea de ensamblaje.

Hay muchos otros ejemplos de conductas específicas a la situación, en las cuales los estudiantes aprenden a dar la respuesta correcta en clase o en una prueba de lápiz y papel, pero la abandonan cuando enfrentan una situación de la vida real en la que podría ser aplicable. Buena parte del aprendizaje que se da en el salón de clase se relaciona con conductas en contextos no académicos, pero la relación está lejos de ser perfecta. La generalización del conocimiento y las habilidades del salón de clases a las situaciones de la vida real es particularmente endeble en el caso del conocimiento verbal. Los maestros se dan cuenta de que si la escuela debe preparar a los estudiantes para la vida, las habilidades y el conocimiento deben enseñarse de tal manera que se transfieran a situaciones laborales y otros contextos no académicos. Los maestros de ciencia, atletismo, arte dramático, música, artes industriales, expresión oral, lenguas extranjeras, caligrafía, agricultura, y muchas otras áreas temáticas, reconocen la necesidad de que los estudiantes practiquen repetidamente y tengan experiencia directa para que las habilidades sean bien aprendidas y transferibles. Los laboratorios y proyectos de ciencia, las habilidades psicomotrices aprendidas en juegos y deportes, tocar instrumentos musicales y cantar, actuar en obras, construir o aplicar objetos útiles en un taller, practicar el hablar en público y la conversación en español y en otros idiomas, todo lo anterior, proporciona oportunidades para aprender y practicar habilidades que son potencialmente útiles fuera de la clase y servirán como cimientos para el aprendizaje práctico experiencial posterior. Debido a un mayor realismo que las pruebas escritas, a las pruebas de ejecución en ocasiones se les conoce como *evaluación auténtica* o, para enfatizar que son una opción a las pruebas escritas, *evaluación alternativa.*

Aunque puede no ser necesario seguir una taxonomía de objetivos psicomotrices al planificar una prueba para medir qué tan bien ha aprendido una persona una habilidad particular, es útil elaborar una lista detallada de las conductas que son indicadoras de un rango de competencia en esa habilidad. Deben tomarse de antemano decisiones como qué tanto peso (numérico) se dará a cada aspecto del desempeño y qué deducciones (si las hay) se harán por errores, lentitud o descuido.

Una prueba de ejecución debe concentrarse, sobre todo, en el producto o resultado final de ejecutar una habilidad, pero también es importante observar la forma en que se realiza (el proceso). Por ejemplo, lo que cuenta más al jugar golf es el número de golpes requeridos para meter la bola en el hoyo, pero todos los instructores de golf se dan cuenta de que la forma, o estilo, también es importante. En las pruebas de ejecución que involucran un producto terminado tangible no sólo debe advertirse la cantidad y calidad del producto, sino también la eficiencia con la que fue hecho.

Tanto los productos como los procesos del desempeño suelen evaluarse de manera subjetiva, principalmente por observación combinada con un registro escrito o electrónico y una lista de verificación o escala de calificación. Es posible examinar y evaluar *portafolios* enteros, o colecciones de los desempeños y productos de los estudiantes a lo largo de un periodo. Para la evaluación precisa del desempeño es crucial una observación cuidadosa que esté tan libre de sesgos como sea posible. Las pruebas de ejecución estructuradas, en las cuales se prueba a cada examinado bajo las mismas condiciones, suelen ser más objetivas que las no estructuradas, donde se observa y evalúa a los estudiantes de manera subrepticia durante la clase, en los pasillos o en otras áreas de la escuela. Pero incluso aunque se tenga sumo cuidado, por su misma naturaleza, las pruebas de ejecución son menos objetivas y, en consecuencia, menos confiables que las pruebas escritas. Además, las pruebas de ejecución requieren más tiempo que las escritas y a menu-

do también equipo costoso y otras condiciones que consumen tiempo. Por esas razones, siempre que el costo y la ineficiencia de una prueba de ejecución no sean compensados por su carácter realista, es preferible una prueba escrita.

RESUMEN

Este capítulo trata principalmente de procedimientos para diseñar y elaborar pruebas de rendimiento educativo, pero los principios analizados también pueden aplicarse a otros tipos de instrumentos de evaluación psicológicos y educativos.

El primer paso en la elaboración de una prueba de rendimiento es preparar una lista de los objetivos conductuales que van a evaluarse. Luego debe construirse una tabla de especificaciones que presente el número de reactivos necesarios en cada categoría de contenido (temático) para cada objetivo conductual. Se han propuesto varias taxonomías o métodos de clasificación de objetivos conductuales en los dominios cognoscitivo, afectivo y psicomotriz. La taxonomía de objetivos educativos más popular es la *Taxonomía de objetivos educativos: el dominio cognoscitivo,* de Bloom y Krathwohl.

Tanto las pruebas de ensayo como las objetivas poseen ventajas y desventajas. Los reactivos de ensayo son más fáciles de elaborar, pero los reactivos objetivos pueden calificarse de manera más rápida y precisa. Las pruebas objetivas también proporcionan una muestra más representativa del contenido de la materia. Las preguntas de respuesta corta, de verdadero y falso, de opción múltiple y de aparejamiento son variedades de los reactivos objetivos. De éstos, los reactivos de opción múltiple son los más versátiles y populares.

Al formar una prueba debe prestarse atención a factores como la longitud y el formato, el método para registrar las respuestas, las facilidades para la reproducción de la prueba, y las instrucciones para la aplicación. Las instrucciones de aplicación de una prueba incluyen el(los) propósito(s), los límites de tiempo, el procedimiento de calificación y lo aconsejable de adivinar cuando se tenga duda.

Las pruebas orales no se usan tan a menudo como las pruebas escritas, pero cuando se planean, aplican y evalúan con cuidado pueden proporcionar información que por lo regular no se obtiene con otros métodos de evaluación. En cierto sentido, tanto las pruebas escritas como las orales son medidas de ejecución, pero el concepto de pruebas de ejecución por lo general se ha concentrado en conducta no verbal. Dado que las pruebas de ejecución son más realistas que las verbales, en ocasiones se les conoce como evaluación auténtica. En lugar de limitarse a describir cómo hacer algo o qué se hizo, las pruebas de ejecución requieren que los examinados demuestren un proceso. Dichas pruebas se emplean de manera extensa para evaluar habilidades aprendidas en el laboratorio y en situaciones de campo, las cuales abarcan desde el laboratorio de ciencia hasta la arena deportiva y otros contextos aplicados. A menudo los maestros conservan y evalúan portafolios del desempeño y los productos de los estudiantes.

PREGUNTAS Y ACTIVIDADES

1. Elija un tema para desarrollar una prueba en un área que le interese, plantee sus objetivos conductuales y de contenido, elabore una tabla de especificaciones y diseñe una prueba objetiva de una hora sobre el tema elegido usando varios tipos de reactivos.

2. Diseñe un sistema de objetivos educativos del dominio cognoscitivo para su salón de clases. ¿En qué difiere de los sistemas que se describieron en el texto? ¿Qué ventajas y desventajas particulares posee?

3. Elabore una taxonomía de objetivos para las humanidades o el currículo básico en la universidad. Incluya al menos cinco objetivos de su taxonomía, con dos o tres subobjetivos bajo los cinco objetivos principales. Defina cada una de las principales categorías y subcategorías de su taxonomía de manera tan clara y objetiva como sea posible.

4. Diseñe una tabla de especificaciones para una prueba global de humanidades que vaya a aplicarse a todos los estudiantes al final de su segundo año en la universidad. Base los objetivos conductuales y de contenido de su tabla de especificaciones en la taxonomía que elaboró en la actividad 3.

5. Describa las fortalezas y debilidades relativas de las pruebas de ensayo, orales y de ejecución. ¿Para qué propósitos y bajo qué condiciones es más apropiado cada tipo de prueba?

6. ¿Por qué suele considerarse que los reactivos de opción múltiple son superiores a los otros tipos de reactivos objetivos? ¿Puede pensar en una situación donde los reactivos de verdadero y falso, completamiento o aparejamiento sean preferibles a los de opción múltiple?

7. Escriba cinco reactivos de respuesta corta (completamiento), cinco de verdadero y falso, y cinco de opción múltiple basándose en la siguiente selección adaptada de Aiken (1980):

> Una razón para la escasez de datos psicométricos sobre los adultos mayores es que en este grupo de edad la gente, cuya conducta es menos susceptible de ser controlada por psicólogos y educadores, a menudo se muestra renuente a ser examinada. Hay muchas razones para explicar la poca cooperación de los adultos mayores en las situaciones de prueba, incluyendo la falta de tiempo, la percepción de las tareas de la prueba como triviales y sin sentido, y el temor de salir mal y parecer tontos. A los adultos mayores, en mayor medida que los adultos más jóvenes que están más conscientes de la prueba, no les entusiasma realizar tareas que los hagan ver ridículos o que son percibidas como irrelevantes en sus vidas.
>
> Debido a que los adultos mayores tienen poca motivación para ser examinados, se requiere sensibilidad y tacto de parte de los examinadores psicológicos para obtener respuestas válidas. Por desgracia, a menudo se cuestiona si los examinadores técnicamente competentes pero jóvenes pueden establecer suficiente *rapport* con los examinados mayores como para comunicarles adecuadamente las instrucciones de la prueba y estimularlos para hacer lo mejor que puedan. Relativamente pocos psicómetras parecen tener el entrenamiento y la experiencia suficientes en la examinación psicológica de los adultos mayores como para hacer un trabajo creíble. Sin embargo, la mayoría de los examinadores encuentra que una vez que las personas mayores aceptan ser probadas, tienen una motivación tan alta como la de los examinados jóvenes para hacer las cosas bien.
>
> Incluso cuando los adultos mayores se muestran cooperativos y motivados, los límites de tiempo de muchas pruebas, la presencia de defectos sensoriales, la tendencia a la distracción y la facilidad con que se fatigan les dificulta desempeñarse de manera satisfactoria. Una de las cosas más características acerca de ser mayor es que los reflejos y los movimientos físicos tienden a ser más lentos. Por esta razón, las explicaciones de la declinación relacionada con la edad en las calificaciones de pruebas en áreas como el aprendizaje y la memoria deben considerar el hecho de que los adultos mayores por lo general no reaccionan con tanta rapidez como los adultos jóvenes.
>
> Aunque la gente mayor suele estar en desventaja en las pruebas cronometradas, su desempeño mejora de modo significativo cuando se le da tiempo suficiente para responder. En las pruebas que no están cronometradas los adultos mayores muestran poca o ninguna inferioridad en comparación con los adultos más jóvenes.

Los defectos sensoriales, en especial en las modalidades visual y auditiva, también pueden interferir con el desempeño en la vejez. Puede ser útil contar con materiales especiales de prueba, como caracteres grandes, y examinadores entrenados que estén alerta en cuanto a la presencia de defectos sensoriales. Sin embargo, en ocasiones un supuesto defecto sensorial en realidad puede ser una máscara para ocultar un problema de lectura o comprensión auditiva. El autor tuvo la experiencia de prepararse para probar a un hombre anciano que, avergonzado por su poca habilidad para la lectura, convenientemente olvidó sus lentes y, en consecuencia, no pudo leer los materiales de la prueba.

8. ¿Cuáles son las ventajas y desventajas de las pruebas orales en comparación con las pruebas escritas? ¿En qué circunstancias son apropiadas las pruebas orales? ¿Cómo deberían diseñarse, aplicarse y calificarse?

9. ¿Qué miden las pruebas de ejecución que no pueda ser medido por medio de pruebas de lápiz y papel (escritas) o por pruebas orales? Describa dos o tres pruebas de ejecución que haya presentado.

CAPÍTULO TRES

ADMINISTRACIÓN, APLICACIÓN Y CALIFICACIÓN DE LOS TESTS

Sin importar qué tan cuidadosamente se elabore una prueba, los resultados no tienen ningún valor si no se administra y califica ésta en forma adecuada. La necesidad de contar con procedimientos y guías establecidos para administrar y calificar pruebas psicológicas y educativas es reconocida por todas las organizaciones profesionales dedicadas a la evaluación de personas. Una fuente importante de estos recursos son los *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association y National Council on Measurement in Education, 1999), una serie de 264 normas para construir, evaluar, administrar y calificar pruebas y otros instrumentos psicométricos, y para interpretar y usar los resultados. Las 16 normas que conciernen específicamente a la administración, calificación y registro de pruebas ponen énfasis en la importancia de tener instrucciones claras para que la administración y calificación se sigan con todo cuidado. Las normas también hacen hincapié en que los materiales de las pruebas deben conservarse seguros, los fraudes han de detectarse y controlarse, y la interpretación de los resultados debe ser clara al leerla.

APLICACIÓN DE LOS TESTS

El procedimiento que debe seguirse para aplicar una prueba o cualquier otro instrumento psicométrico depende del tipo de que se trate (individual o colectiva, con tiempo predeterminado o sin éste, cognoscitiva o afectiva), lo mismo que de la edad cronológica, la educación, los antecedentes culturales y el estado físico y mental de los examinados. Cualesquiera que sean el tipo de prueba y las características de quienes se someten a ella, el desempeño también puede alterarse por factores como disposición y motivación del examinado, cantidad de sueño durante la noche previa a la prueba, molestias físicas, angustia relativa a la prueba, otros problemas emocionales, y medicamentos que se estén consumiendo.

No sólo la disposición, la habilidad para resolver pruebas y la motivación de los examinados afectan el desempeño, sino también la apariencia y el comportamiento de quien aplica la evaluación, así como la situación. Sobre todo, en el caso de pruebas individuales, son importantes la habilidad y la personalidad del examinador. Quienes administran la mayoría de las pruebas individuales deben tener un título o certificado formal expedido por un organismo gubernamental apropiado o ser supervisados por otro examinador certificado. Estos requisitos contribuyen a garantizar que los examinadores cuenten con el conocimiento y la capacidad necesarios para administrar, calificar e interpretar diversos tipos de instrumentos psicométricos.

Las variables situacionales, incluyendo el tiempo para resolver la prueba y el lugar donde se aplique, y condiciones ambientales como iluminación, temperatura, nivel de ruido, ventilación u otras distracciones, también pueden contribuir a la motivación, concentración y desempeño de las personas que se examinan. Por consiguiente, antes de administrar una prueba, debemos estar seguros de que el ambiente físico sea el apropiado.

Deberes del examinador antes de la prueba

Programación. Al programar una prueba, el examinador debe tomar en cuenta las actividades que suelen realizar los examinados en esa hora del día. No es sensato administrar pruebas a niños durante las horas del almuerzo o del juego, cuando acostumbran realizar alguna otra actividad placentera, o cuando acaban de tener lugar acontecimientos divertidos o emocionantes (por ejemplo, inmediatamente después de días feriados). El tiempo de la prueba casi nunca debe excederse de una hora al tratarse de niños pequeños o de una hora y media cuando son niños de secundaria. Debido a que 30 minutos es el límite de tiempo en que un niño de nivel preescolar y de primaria puede permanecer atento a las tareas de una prueba, puede requerirse más de una sesión para administrarse pruebas extensas a niños pequeños.

Con respecto a las pruebas en el aula, debe informarse a los estudiantes con suficiente anticipación cuándo y dónde se administrará la prueba, qué contenido de materias incluirá, qué tipo de prueba (objetiva, de ensayo, oral) se administrará y cuánto tiempo se concederá para resolverla. Los estudiantes merecen la oportunidad de prepararse intelectual, emocional y físicamente para una prueba. Por ello, regularmente no es aconsejable imponer exámenes sorpresa. Si el maestro piensa que ocasionalmente las pruebas sin previo aviso ayudan a garantizar que los alumnos se mantengan al corriente con el material del curso, dichos exámenes no deben tener el mismo peso que las evaluaciones habituales.

Consentimiento informado. En muchos lugares, la aplicación de una prueba de inteligencia o de otro instrumento de psicodiagnóstico a un niño requiere del consentimiento informado de uno de los padres, un tutor o de otra persona legalmente responsable del niño. El *consentimiento informado* consiste en un acuerdo entre una institución o individuo y una persona en particular o su representante legal. Con los términos del acuerdo se otorga permiso para aplicar tests psicológicos a una persona y/o conseguir otra información con propósitos de evaluación o de diagnóstico.

> Debe obtenerse el consentimiento informado de quienes se someterán a una prueba, o de sus representantes legales cuando sea adecuado, antes de iniciarla excepto (a) cuando la evaluación sin consentimiento sea ordenada por ley o por reglamentación gubernamental; (b) cuando la evaluación sea parte de las actividades habituales de la escuela, o (c) cuando el consentimiento esté claramente implícito (American Educational Research Association *et al*., 1999, p. 87).

El requisito de consentimiento informado suele cumplirse al obtener la firma de una persona legalmente responsable en una forma estándar proporcionada por el distrito escolar u otra institución pertinente. La forma especifica el(los) objetivo(s) de la evaluación, el uso que se hará de los resultados, los derechos del padre o tutor y el procedimiento a seguir para obtener una copia del informe final o de la interpretación.

Familiarizarse con la prueba. No debe haber duda en cuanto a la familiaridad con el material de la prueba y el procedimiento de aplicación cuando el examinador es la misma persona.

Debido a que la persona que administra una prueba estandarizada rara vez es la misma que la elaboró, debe estudiarse con cuidado el manual adjunto antes de iniciar el proceso de evaluación. Es de particular importancia familiarizarse con las instrucciones de administración y con el contenido de la prueba. Para lograr esta familiaridad, es recomendable que el examinador mismo se someta a la prueba antes de administrarla a otra persona. Por último, es aconsejable revisar las instrucciones y otros materiales del procedimiento justo antes de la aplicación. Asimismo, los folletos, las hojas de respuestas y otros materiales de la prueba deben revisarse y contarse con anterioridad. Las *pruebas seguras* que incluyen un número de serie, como la Prueba de Evaluación Académica y los Exámenes de Registro de Graduados, deben inspeccionarse en forma minuciosa y ordenarse por número.

Cuando un niño o un adulto sea remitido para un examen psicológico por una institución externa o por un médico o juez, las pruebas y otros procedimientos de psicodiagnóstico a administrarle dependerán de los tipos de información que requiera la fuente que remite y de los fines para los que se empleará la prueba. En consecuencia, es importante que la persona que remite especifique con precisión la información requerida y lo que se hará con ella. En todo caso, el examinador debe estar familiarizado a fondo con las pruebas o los demás instrumentos psicométricos y los tipos de individuos y condiciones para los que son adecuados.

Garantizar condiciones de evaluación satisfactorias. Los examinadores deben asegurarse de que los asientos, la ventilación, la temperatura, el nivel de ruido y otras condiciones físicas del ambiente de evaluación sean satisfactorios. Es preferible utilizar una habitación que sea familiar para los examinados y esté relativamente libre de distracciones. Colocar un letrero de "Prueba-No molestar" en la puerta cerrada puede contribuir a evitar interrupciones y otros distractores. También deberá contarse con acceso fácil a las salidas y a las instalaciones sanitarias.

Es mejor administrar una prueba individual en una habitación privada, sólo con el examinador, el examinado y, de ser necesario, uno de sus padres, el tutor u otra persona responsable. Ya sea en una prueba individual o en una colectiva, es preciso tomar previsiones especiales para examinados físicamente discapacitados o diferentes (por ejemplo zurdos).

Reducir los fraudes. Los examinadores bien capacitados están muy conscientes de la importancia de la seguridad de la prueba, tanto antes como después de administrarla, y de aceptar la responsabilidad de que se conserve dicha seguridad.

> Debe advertirse a quienes se someten a una prueba que hacer que otra persona responda el examen en su lugar, revelar material confidencial o cualquier otra forma de fraude es un comportamiento inadecuado que puede generar sanciones (American Educational Research Association *et al*., 1999, p. 88).

Antes de la prueba, debe procurarse que haya asientos confortables que además contribuyan a eliminar la posibilidad de fraude. Aunque es preferible, no siempre puede lograrse que los examinados dejen un asiento libre entre cada uno para que resulten difíciles las oportunidades de hacer trampa. Preparar formas múltiples (con reactivos distintos o con una distribución diferente) de la prueba y repartir formas distintas a los examinados adyacentes puede reducir las trampas en una prueba aplicada colectivamente. Otra posibilidad es usar diversas hojas de respuestas, es decir, con distinta disposición. También deben emplearse varios vigilantes cuando se trate de pruebas para un grupo grande. Ellos pueden ayudar a distribuir y recoger los materiales de la prueba y a responder dudas sobre el procedimiento; además, su presencia tiende a desalentar las conductas tramposas y la indisciplina. La vigilancia y otros procedimientos diseñados para pre-

venir las trampas se consideran con suma seriedad al administrar pruebas estandarizadas seguras, tales como la Prueba de Evaluación Académica y los Exámenes de Registro de Graduados. Estas pruebas, los folletos y las hojas de respuestas, que se cuentan con todo cuidado antes y después de los exámenes, se supervisan minuciosamente. Se solicita de las personas que se someten a estas pruebas mostrar una identificación oficial antes de ser admitidos en el aula de exámenes.

Deberes del examinador durante la prueba

Seguir las instrucciones de la prueba. Mediante instrucciones meticulosamente preparadas, que se leen en forma lenta y clara al presentarse oralmente, se informa a los examinados sobre los objetivos de la prueba y cómo anotar sus respuestas. Se pide a los examinadores de pruebas estandarizadas que sigan cuidadosamente las instrucciones de administración, aun cuando una explicación adicional podría esclarecer la tarea para los examinados. El no apegarse a las instrucciones estándar puede dar como resultado una tarea distinta de la que tenían en mente los diseñadores de la prueba. Si las instrucciones no son idénticas a las presentadas a la muestra de personas con las que se estandarizó la prueba, los resultados no tendrán el mismo significado que los del grupo de estandarización. Por ende, se habrá perdido un útil marco de referencia para interpretar los resultados.

Los examinadores de contextos clínicos y educativos en ocasiones van más allá de las instrucciones de la prueba e intentan probar los límites de las habilidades o las características personales de los examinados. Esto puede lograrse mediante procedimientos de evaluación *dinámicos* o *auténticos* para obtener claves adicionales con fines de interpretación o diagnóstico. Una ilustración de la evaluación dinámica se encuentra en el concepto de Feuerstein acerca de la *evaluación del potencial de aprendizaje* (Feuerstein, Feuerstein y Gross, 1997). La evaluación del potencial de aprendizaje implica un formato de prueba-enseñanza-prueba donde se examina a una persona, se le somete a una práctica sobre los materiales de la prueba y luego vuelve a aplicársele el examen. Se calcula entonces el cambio en el nivel de desempeño de la primera a la segunda ocasión en que se resuelve la prueba como una medida del potencial de aprendizaje del examinado (vea también Tombari y Borich, 1999).

Permanecer alerta. Al administrar una prueba colectivamente, ya sea estandarizada o no, el examinador debe mantenerse alerta para evitar las trampas, así como que se hable o haya ruido innecesario. También es sensato tomar la precaución de tener un mensajero disponible para casos de emergencias médicas o algún otro problema. En pruebas elaboradas por el maestro, o incluso en pruebas estandarizadas si las instrucciones lo permiten, es posible informar periódicamente a los estudiantes cuánto tiempo les resta anotando la hora en el pizarrón o en otra superficie visible.

Establecer una relación interpersonal. Tanto en pruebas individuales como colectivas, el comportamiento del examinador puede tener un efecto considerable en la motivación y el comportamiento de los examinados. En ocasiones, hasta una sonrisa puede brindar ánimo a los examinados nerviosos o poco preparados a fin de que conserven la calma y logren un desempeño óptimo. Debido a que las pruebas individuales proporcionan una mejor oportunidad de observar a los examinados que las pruebas de aplicación colectiva, es más probable detectar falta de motivación, distracción y tensión en un contexto de evaluación individual. Así, pueden realizarse esfuerzos para manejar estos problemas o por lo menos tomarlos en cuenta al interpretar los resultados. En una situación de evaluación colectiva, donde suele ser imposible mantener una interacción personal con cada uno de los examinados, el examinador tiene más dificultades para

advertir cómo se está sintiendo y desempeñando una persona. Tanto en pruebas individuales como colectivas, una buena regla a seguir es mostrarse amigable pero objetivo, con autoridad mas no autoritario, con modales y vestuario apropiados y a cargo de la situación de evaluación. Tal comportamiento por parte del examinador tiende a crear una situación de *rapport*, es decir, una relación interpersonal cordial de aceptación que anima a los examinados a responder en forma honesta y precisa.

Prepararse para manejar problemas especiales. En determinadas circunstancias, los examinadores deben ser especialmente activos y alentadores. Una situación de evaluación produce cierta cantidad de tensión casi en cualquier persona, y en ocasiones un examinado se pone muy nervioso. Las pruebas en personas muy jóvenes, muy viejas, perturbadas mentalmente, con retraso mental, discapacidad física o desventajas culturales presentan problemas especiales. En algunas situaciones, tal vez tengan que darse las preguntas y las respuestas en forma oral y no escrita, o en una lengua en particular. El examinador no sólo debe estar familiarizado con el material de la prueba, sino también conducirse de manera alerta, flexible, cálida y objetiva. Estas cualidades no se enseñan con facilidad, pero la experiencia en diversas situaciones de evaluación desempeña un papel importante para adquirirlas.

Flexibilidad. También se permite cierta flexibilidad al administrar pruebas no estandarizadas e incluso en algunos instrumentos estandarizados, pero el exceso de flexibilidad puede volver inútiles las normas establecidas con propósitos de interpretación. Al evaluar con estas medidas, la sensibilidad y la paciencia por parte del examinador pueden proporcionar una mejor oportunidad para que los discapacitados y otros individuos con problemas especiales demuestren sus aptitudes. Otros procedimientos recomendados, que se han adaptado de técnicas de instrucción reconocidas, son los siguientes:

1. Proporcionar tiempo suficiente para que los examinados respondan el material de la prueba.
2. Permitir la práctica necesaria con reactivos de ejemplo.
3. Usar periodos de evaluación relativamente cortos.
4. Observar si hay signos de fatiga o angustia y tomarlos en cuenta.
5. Ser consciente y tomar las medidas pertinentes en caso de defectos visuales, de audición y otros sensoriales o perceptuales-motrices.
6. Brindar estímulo y refuerzo positivo con generosidad.
7. No intentar obligar a los examinados a responder cuando se han resistido a hacerlo en varias ocasiones.

Pruebas orales. Los exámenes orales a menudo provocan en los estudiantes sentimientos encontrados y mucha aprehensión. Como resultado, los esfuerzos por calmar esos temores y ofrecer otros métodos de evaluación a quienes les afectan emocionalmente las situaciones de evaluación oral puede mejorar la efectividad de este tipo de pruebas. Los examinadores que realizan esfuerzos especiales por establecer una relación interpersonal con los examinados descubren que es posible que éstos lleguen a disfrutar de las pruebas orales.

Aplicar una prueba

En general, no se consideran justos los exámenes sorpresa. Los alumnos merecen la oportunidad de prepararse para una prueba. Debe informárseles con anticipación no sólo cuándo y dónde se realizará la prueba, sino también lo que abarcará y qué tipo de prueba será. Con respecto al for-

mato, tanto los estudios en aulas como de laboratorio han revelado que las personas tienden a obtener mejores resultados en pruebas de recordatorio (ensayo, de respuestas breves) cuando se les informa que se administrará una prueba de ensayo (por ejemplo, May y Thompson, 1989). Esperar una prueba de reconocimiento (opción múltiple, de falso-verdadero) estimula un estudio de los detalles más concentrado, mientras que esperar una prueba de recordatorio origina mayores esfuerzos por recordar unidades de un nivel superior y temas del material (Schmidt, 1983).

Hay estudios de aula donde se ha descubierto que comunicar con anticipación que se administrará una prueba objetiva está relacionado con calificaciones más elevadas en pruebas de opción múltiple, de falso-verdadero y otras de reconocimiento. Sin embargo, los resultados de estudios de laboratorio son más complejos (Lundeberg y Fox, 1991). Además, otros factores como la habilidad mental, la habilidad para resolver pruebas, adivinar y una cuidadosa lectura y consideración de los reactivos parecen tener tanto efecto en las calificaciones de las pruebas como saber qué tipo de prueba se aplicará. En cualquier caso, al aplicar una prueba de aprovechamiento, es justo proporcionar información previa sobre su forma y cobertura.

Ingenio para resolver pruebas. Al responder reactivos de prueba objetivos, la gente suele emplear métodos muy diferentes de los que previó el autor de los reactivos. No todos los examinados leen con cuidado los reactivos y con frecuencia no utilizan la información proporcionada. Esto puede no ser esencial en todos los casos, ya que en ocasiones es posible reconocer las respuestas correctas en reactivos de opción múltiple sin haber leído el material en que se basan las preguntas. Por ejemplo, las opciones erróneas pueden descartarse al advertir que algunas están expresadas en forma incorrecta o son demasiado extensas o breves. Otras claves que pueden revelar las respuestas correctas en reactivos de opción múltiple son asociaciones aliteradas, opciones no relacionadas, lenguaje incluyente, opciones en clave que son más precisas que las demás, claves gramaticales y opciones que se revelan al aparecer resueltas en otros reactivos.

Las observaciones de estudiantes que responden pruebas de opción múltiple y luego son entrevistados revelan que, si bien los reactivos a menudo se responden simplemente eliminando las opciones que parecen incorrectas, una práctica más común es realizar juicios comparativos entre las opciones. Los resultados de la investigación de Rogers y Yang (1997) indican que los alumnos primero deben tener cierto conocimiento del contenido de las opciones raíz y/o de los reactivos a fin de eliminar las opciones incorrectas y aprovechar las claves del reactivo.

Otro aspecto del ingenio para resolver pruebas es el conocimiento de la idiosincracia del maestro. *El ingenio para resolver pruebas* parece ser una capacidad específica, no general, para identificar claves que se desarrolla en los estudiantes al madurar y compartir información sobre la forma de resolver pruebas (Evans, 1984). Por ejemplo, la extensión, el tecnicismo y cierto exotismo de las opciones proporcionan claves para encontrar las respuestas correctas (Strang, 1980; Tidwell, 1980). También es de interés el hecho de que la influencia del ingenio para resolver pruebas en general es mayor para los reactivos de cuatro opciones que para los de tres (Rogers y Harley, 1999). Los niños parecen tener más esta habilidad que las niñas (Preston, 1964), y los reactivos verbales son más susceptibles de resolverse mediante esta capacidad que los numéricos (Rowley, 1974). Algunos aspectos de dicha habilidad pueden enseñarse (American College, 1978; Millman y Pauk, 1969). El cuadro 3.1 contiene una lista de 15 sugerencias que, al practicarse antes y durante una prueba, pueden incrementar el ingenio para responder pruebas y mejorar los resultados.

Cambiar las respuestas. A menudo los examinados se enfrentan a la duda de cambiar o no sus respuestas iniciales a los reactivos. En ocasiones se afirma que, como las primeras respuestas suelen ser las correctas, revisar la prueba y cambiar las respuestas sobre las que ya se ha re-

CUADRO 3.1
SUGERENCIAS PARA MEJORAR LOS RESULTADOS DE SU PRUEBA

Antes de la prueba

1. Pida al instructor una copia de viejas pruebas que pueda usted revisar legítimamente.
2. Pregunte a otros estudiantes qué tipo de pruebas suele administrar el instructor.
3. No espere a que llegue el día anterior para empezar a estudiar cuando la prueba ha sido anunciada con anticipación.
4. Estudie para el tipo de prueba (de elección múltiple, de falso-verdadero, de ensayo) que se ha anunciado.
5. Si no se ha especificado el tipo de prueba que se administrará, tal vez sea mejor estudiar para una prueba de recordatorio (ensayo).
6. No convierta el estudiar para una prueba en un acontecimiento social; en general es mejor aislarse para preparar una prueba.
7. No se ponga demasiado cómodo para estudiar. Su cuerpo supone que desea dormir cuando se recuesta o su posición resulta demasiado confortable.
8. Intente estructurar el material que estudia como reactivos de prueba, por ejemplo, en reactivos de opción múltiple si éste es el tipo de examen que tendrá, o en reactivos de ensayo si está programada una prueba de ensayo.
9. Aplique el Estudio Q3R (inspección, preguntas, lectura, recitación, revisión) al estudiar para una prueba. Revise el material, formúlese preguntas acerca de él, lea con atención intentando recordar, recite el material para usted mismo después de leerlo y revíselo justo antes de la prueba.

Durante la prueba

1. Lea cuidadosamente las instrucciones de la prueba antes de empezar con las preguntas. Si cierta información, como los límites de tiempo, la corrección por adivinar, el peso de los reactivos o cuestiones similares se han omitido, no dude en preguntar al examinador.
2. En pruebas de ensayo, piense en las preguntas y formule respuestas en su mente y/o en un papel aparte antes de empezar a escribir las respuestas definitivas.
3. Tómese su tiempo al resolver una prueba. Por ejemplo, en una prueba de opción múltiple debe haber respondido una n fracción de la prueba para cuando haya transcurrido la *n* fracción del tiempo reglamentario.
4. Ya sea que se emplee o no la corrección por adivinar al calificar una prueba, no deje reactivos sin responder si puede descartar por lo menos una opción.
5. Pase por alto los reactivos más difíciles y regrese a ellos posteriormente. No entre en pánico si no puede responder un reactivo; enciérrelo en un círculo y regrese a él después de haber resuelto otros. Entonces, si aún no está seguro de la respuesta, reflexione y adivine la mejor opción.
6. No se apresure a entregar la prueba antes de que termine el tiempo; cuando le sea posible, revise sus respuestas.

flexionado es contraproducente (Benjamin, Cavell y Shallenberger, 1984). Sin embargo, los resultados de varias investigaciones indican que los examinados tienden a obtener calificaciones más altas cuando reconsideran sus respuestas y cambian aquellas sobre las que dudan (vea, por ejemplo, Geiger, 1990, 1991a, 1991b). Es más probable que las respuestas erróneas se conviertan en acertadas que viceversa, aunque la cantidad de preguntas que de hecho se cambian tiende a ser pequeña.

Adivinar. Las instrucciones para pruebas objetivas suelen incluir consejos sobre omitir un reactivo o adivinar cuando se duda sobre la respuesta correcta. Adivinar, lo que es más probable cuando los reactivos son difíciles o muy elaborados, origina más inflación de la calificación en reactivos de falso-verdadero que en pruebas de opción múltiple. En general, es aconsejable adivinar sólo cuando pueden eliminarse una o más opciones, o cuando se tiene cierta idea sobre la opción correcta. Debido a que en general es posible eliminar por lo menos una opción en un reactivo, adivinar antes que omitir reactivos suele producir calificaciones más altas. Esto es cierto ya sea que se "corrijan" o no los resultados por adivinar.

Como es comprensible, los examinados adivinarán menos si se les informa que su calificación será reducida como castigo por adivinar, al contrario de lo que sucede cuando no se dan instrucciones al respecto o se les pide que adivinen cuando tengan dudas. Desafortunadamente, los estudiantes no siempre leen ni siguen con atención las instrucciones. Incluso quienes las leen palabra por palabra no siempre las interpretan del mismo modo. Sin importar lo que aconsejen o no las instrucciones, a algunas personas no les gusta tomar riesgos y se muestran renuentes a adivinar cuando no están seguros de la respuesta correcta.

Deberes del examinador después de la prueba

Después de administrar una prueba individual, el examinador debe recoger y guardar en lugar seguro todos los materiales de la prueba. Es preciso animar a los examinados sobre su desempeño, tal vez darles alguna pequeña recompensa si se trata de niños y conducirlos al lugar adecuado. En evaluaciones clínicas, en general es importante consultar con el padre o acompañantes del examinado, quizás antes y después de la prueba. Al terminar la prueba, también se dará información sobre lo que se hará con los resultados a los examinados y/o a sus acompañantes. El examinador tranquiliza a los interesados al prometer comunicarles los resultados e interpretaciones a los propios individuos o a la institución y recomendar medidas subsiguientes.

Tras la administración de una prueba colectiva de grupo, el examinador tiene que recoger los materiales pertinentes (folletos, hojas de respuestas, papel para usar como borrador, lápices, etcétera). En caso de una prueba estandarizada, es necesario contar y cotejar los folletos y las hojas de respuestas, así como verificar todos los demás materiales para asegurarse de que nada falte. Sólo entonces se despide a los examinados o se les prepara para la siguiente actividad, y se ordenan las hojas de respuestas para calificarlas.

Pruebas adaptativas

Históricamente, no se ha seguido con precisión en todas las pruebas un procedimiento de aplicación de pruebas en el que se presenten los mismos reactivos a todos los examinados. No obstante, en general se ha permitido poca flexibilidad al determinar los reactivos. Este método tradicional de aplicación de pruebas es particularmente ineficaz en pruebas de aprovechamiento, porque se les presentan a los examinados muchos reactivos que resultan demasiado fáciles o difíciles para ellos. Adaptar el contenido de una prueba al nivel de capacidad del examinado elimina la necesidad de aplicar muchos reactivos muy fáciles o muy difíciles, lo que ahorra tiempo y esfuerzo.

En pruebas *adaptativas* o *a la medida*, los reactivos específicos aplicados a una persona en particular dependen de su capacidad calculada a partir de sus respuestas a reactivos previos. Debido a que las pruebas son más precisas para medir la habilidad de las personas si la dificultad de los reactivos corresponde a su propio nivel de habilidad, evaluar ésta mientras el exami-

nado avanza en la prueba permite seleccionar los reactivos más cercanos a su habilidad real (vea Meijer y Nering, 1999; Wainer, 2000).

Los bancos de reactivos para pruebas adaptativas pueden ser recopilados por computadoras programadas para seguir alguna de las metodologías de respuesta de reactivos (vea los capítulos 4 y 5). En las pruebas adaptativas, deben cumplirse algunos supuestos de la *teoría de respuesta al Ítem* (*IRT*) incluyendo los siguientes: (1) todos los reactivos de una reserva miden una sola habilidad o dimensión de aprovechamiento, y (2) los reactivos son independientes, es decir, la respuesta de una persona a un reactivo no depende de su respuesta a ningún otro reactivo. El cumplimiento de la primera suposición, de unidimensionalidad, es más probable en el caso de bancos de reactivos o de pruebas derivadas del análisis factorial (vea el apéndice A). La segunda suposición se cumple si los reactivos no están interconectados o interrelacionados de alguna manera.

El procedimiento adaptativo para aplicar una prueba de aprovechamiento o de capacidad funciona de la siguiente manera. Aplicando un modelo estadístico apropiado y una metodología de respuesta de reactivos, se recopila por computadora un banco de reactivos de prueba variando la dificultad y quizás otras características. Una estimación del nivel de habilidad del examinado determina los reactivos que se administrarán primero. Como alternativa, al principio pueden administrarse reactivos de mediana dificultad. La selección de los reactivos que se administrarán subsecuentemente depende de las respuestas del examinado a los reactivos previos. La evaluación continúa mientras el cálculo de error o el nivel de precisión de las respuestas no llegue a un nivel especificado.

A diferencia del procedimiento de evaluación tradicional, en las pruebas adaptativas no se permite a los examinados saltarse reactivos ni revisar o cambiar sus respuestas.[1] Pero debido a que no todos los reactivos de un banco se administran a cada examinado, las pruebas adaptativas son más eficientes que las convencionales. Se aplica al examinado sólo alrededor de la mitad de los reactivos usados en las evaluaciones tradicionales, sin que se pierda información y manteniendo confiabilidad y validez equivalentes.

La calificación de una persona en la mayoría de las pruebas adaptativas está determinada no sólo por el procedimiento tradicional de contar la cantidad de reactivos contestados correctamente, sino tomando en cuenta las características estadísticas de los reactivos. En todo caso, la investigación ha demostrado que las calificaciones de las pruebas adaptativas computarizadas son sumamente comparables a las calificaciones de las pruebas de lápiz y papel equivalentes (Kapes y Vansickle, 1992; Mead y Drasgow, 1992). Asimismo, al administrar reactivos que son más apropiados para el nivel de habilidad del examinado, una prueba adaptativa puede resultar más confiable que otra más extensa diseñada para evaluar la misma habilidad.

La seguridad de una prueba es más fácil de mantener en el caso de pruebas adaptativas asistidas por computadora. El requisito de seguridad es de particular importancia en el caso de sistemas de evaluación proporcionados a través de Internet, como el desarrollado por Northwest Evaluation Association (Olson, 2000). Otras ventajas de las pruebas adaptativas asistidas por computadora incluyen calificaciones y registros más precisos e inmediatos, menos errores generados por adivinación, así como la posibilidad de grabar las respuestas y los tiempos en que se resuelven los reactivos(Bunderson, Inouye y Olsen, 1989). Una desventaja, al menos cuando se

[1]Rocklin, O´Donnell y Holst (1995) propusieron una variante de la evaluación adaptativa computarizada, llamada *evaluación autoadaptativa*, que ofrece a los examinados la oportunidad de diseñar dinámicamente la dificultad de los reactivos y, por tanto, ampliar su estado afectivo y motivacional. En este procedimiento, antes de la presentación, los reactivos en una *prueba autoadaptativa* se agrupan por nivel de dificultad con base en datos normativos. Se permite al examinado especificar la categoría de dificultad de la que debe tomarse cada reactivo sucesivo. De esta manera, un examinado que busque un reto puede especificar que el siguiente reactivo sea difícil, mientras otro que esté intentando evitar el fracaso puede especificar que el siguiente reactivo sea bastante fácil.

evalúan individuos o grupos pequeños, es el costo de la inversión inicial y el gasto por el mantenimiento del equipo y de la actualización de los programas de cómputo.

Los usos de las pruebas adaptativas para evaluar la inteligencia general y las habilidades eran más bien limitados hasta hace relativamente poco. Algunos organismos ofrecen versiones adaptativas computarizadas de la Prueba de Evaluación Académica (SAT), la Batería de Aptitudes Vocacionales de las Fuerzas Armadas (ASVAB), los Exámenes de Registro de Graduados (GRE) y algunas otras pruebas de habilidades cognoscitivas (vea Bergstrom y Lunz, 1999; Mills, 1999; Segall y Moreno, 1999) y de personalidad (vea, por ejemplo, Forbey, Handel y Ben-Porath, 2000; Reise y Henson, 2000).

CALIFICACIÓN DE LOS TESTS

Los diseñadores profesionales de tests no esperan a que se elabore y administre una prueba antes de decidir qué procedimiento de calificación usarán. En una prueba realizada por maestros consistente en varias partes que incluyen distintos contenidos o tipos de reactivos, es posible que el maestro quiera obtener calificaciones separadas de las diversas partes, así como un resultado general de la prueba en su conjunto. También debe decidirse si se restará una corrección por adivinar, si se asignarán distintos pesos a la calificación de los diversos reactivos o secciones y si se entregarán los resultados en forma directa o convertidos a otras escalas numéricas. Para pruebas estandarizadas, el maestro de aula no necesita tomar todas estas decisiones. Las hojas de respuestas pueden calificarse a máquina, y aun cuando se califiquen a mano, pueden usarse plantillas de calificación, proporcionadas por el editor de las pruebas, de acuerdo con las instrucciones incluidas en el manual.

Calificación de pruebas de ensayo

Las pruebas de ensayo pueden hacerse más efectivas al estructurar la tarea con claridad, de modo que la interpretación de una pregunta no varíe mucho de una persona a otra. La calificación puede basarse entonces en la calidad de la respuesta. De manera similar, el procedimiento de calificación para los reactivos de ensayo deberá estar tan estructurado y ser tan objetivo como sea posible, de forma tal que las calificaciones dependan menos de factores ajenos al contenido o de impresiones y más del nivel de conocimiento y comprensión demostrado. Calificar con base en la destreza caligráfica más que en la calidad de las respuestas,[2] generalizar demasiado (*error de indulgencia*) y asignar una calificación alta a una respuesta simplemente porque el examinado obtuvo una buena calificación en otros reactivos (*efecto de halo*), se encuentran entre los errores que pueden alterar las calificaciones en reactivos de ensayo.

Pueden tomarse varias medidas para que las calificaciones de las pruebas de ensayo sean más objetivas y confiables. Para empezar, el examinador debe decidir ya sea calificar la pregunta en conjunto o asignar pesos distintos a cada componente. La calificación completa (*global* u *holística*) es común, pero tal vez sea más significativo emplear un procedimiento *analítico* en el que se asignen puntos para cada reactivo de información o habilidad incluido en la respuesta. En el primer reactivo de ensayo de la tabla 2.4, por ejemplo, puede otorgarse un punto por cada ventaja o desventaja correcta registrada y un máximo de cinco puntos para la forma en que se orga-

[2]Las calificaciones en las pruebas de ensayo no siempre se relacionan positivamente con la calidad caligráfica. En un estudio de Chase (1990-1991), los ensayos escritos con mala letra reciben calificaciones más elevadas que los escritos con buena letra. Esto puede deberse a que los mejores alumnos tenían mala caligrafía, a que los profesores interpretaban la buena letra como un intento por enmascarar el conocimiento deficiente del material, o a algún otro factor.

niza la respuesta. La cantidad máxima de puntos asignados a un reactivo debe determinarse no sólo por el juicio del examinador sobre la importancia del reactivo, sino también por la extensión asignada a la respuesta. Cuando las instrucciones especifican una respuesta de media página, el reactivo debe tener menor peso que cuando se requiere una respuesta de página completa.

Cualesquiera que sean los pesos de calificación asignados a preguntas y respuestas específicas, es aconsejable que el diseñador de la prueba prepare de antemano respuestas ideales para las preguntas. También se recomienda que se bloqueen los nombres de los examinados antes de revisar las pruebas, de modo que puedan evaluarse en forma anónima. A continuación se presentan otras recomendaciones:

1. Califique todas las respuestas a una pregunta antes de pasar a la siguiente pregunta.
2. Califique todas las respuestas a una pregunta específica durante el mismo periodo de calificación.
3. Cuando se califiquen tanto el estilo (mecánica, calidad de la redacción) como el contenido, es preciso evaluar las pruebas en forma separada.
4. Pida a otra persona que califique nuevamente cada prueba y como resultado final elija el promedio de puntos asignados por ella y por usted.
5. Escriba comentarios al lado de las respuestas de los examinados y marque las correcciones en las pruebas.

Las correcciones y los comentarios escritos en las pruebas de aula son un complemento valioso de la cantidad de puntos o del grado asignado. Es más probable que el alumno aprenda algo extra si sus respuestas se corrigen y comentan que cuando sólo se les asigna un número o letra.

Los programas de cómputo para calificar ciertos tipos de reactivos de ensayo ya están disponibles para usarse vía red en un formato de escribir-evaluar-reescribir. Dos ejemplos son Intelligent Essay Assessor, basado en el análisis semántico latente (Landauer, 1998, 1999) y el programa "e-rater". Los procedimientos de calificación y asignación de grados basados en la computadora para los ensayos empiezan por "enseñar" al programa sobre el tema asignado mediante la lectura de cientos de miles de vocablos de textos en línea. Los ensayos escritos por expertos sobre el tema y los ensayos de estudiantes ya calificados por instructores son digeridos por el programa para establecer sus procedimientos de evaluación. Los programas van más allá de verificar simplemente la extensión y mecánica de las palabras y de evaluar el aprendizaje específico de temas y preguntas. Se determinan y evalúan la inclusión de conceptos clave, la estructura semántica y la orientación de los argumentos del alumno. Los programas disponibles en la actualidad para asignar grados a ensayos no determinan la medida en que la escritura resulta creativa o compleja, sino más bien están orientados hacia ensayos que exponen temas objetivos (Murray, 1998).

Calificación de pruebas objetivas

Una ventaja exclusiva de las pruebas objetivas es la eficiencia y precisión con que pueden calificarse. Mientras quienes evalúan pruebas de ensayo dedican en general horas para leer las respuestas y revisar su corrección, un empleado puede calificar una prueba objetiva de manera rápida y precisa mediante una plantilla o una máquina. De modo que es posible regresar las pruebas a los estudiantes cuando aún tienen en mente el material visto en ellas.

Es posible preparar con gran facilidad tiras de claves o plantillas para calificar a mano los cuadernillos de prueba o las hojas de respuestas. Puede elaborarse una tira de claves funcional usando una tira de cartón donde las respuestas correctas se ubiquen en los sitios que corresponden a los espacios de la prueba donde se escriben las respuestas. Para preparar una plantilla de

calificación a usar en hojas de respuestas especiales, en una hoja en blanco o cartoncillo se perfora en los espacios correspondientes a las respuestas correctas.

Calificaciones a máquina. Aunque la mayoría de las hojas de respuestas para pruebas distribuidas comercialmente pueden calificarse a mano o a máquina, las que distribuyen ciertas organizaciones sólo se califican a máquina. Después de que se aplica una prueba, las hojas de respuestas se envían por correo a un servicio especial o se regresan al distribuidor para que las califiquen a máquina.

Las máquinas para calificar han estado disponibles desde la primera mitad del siglo XX. Las máquinas de antaño eran sensibles sólo a marcas magnéticas colocadas en el papel, por lo que se requerían lápices magnéticos especiales para marcar las hojas de respuestas. Las máquinas contemporáneas para calificar lotes grandes de hojas de respuestas son lectores ópticos sensibles a las marcas realizadas con lápices comunes.

No se requiere de una computadora para la calificación rápida y eficiente de pruebas, pero utilizarla provee de cierta flexibilidad de uso que posteriormente permite realizar análisis estadísticos, y la interpretación y almacenamiento de las calificaciones y otros datos personales. Además de la calificación realizada localmente con un lector óptico, las hojas de respuestas pueden enviarse por correo o módem a un servicio de calificación central.

La programación requerida para usar un lector óptico de escritorio es bastante sencilla e incluye un rango amplio de funciones, tales como ponderar reactivos, calificar parcialmente, analizar reactivos, marcar aciertos y errores e imprimir diversos tipos de información, estadísticas y gráficas. Adicionalmente a las calificaciones sin depurar y a las convertidas, se registran distribuciones de frecuencia e histogramas, estadísticas de pruebas (medias aritméticas, desviaciones estándar, coeficientes de consistencia interna) y estadísticas de reactivos (índices de dificultad y de discriminación, distribución de las respuestas a opciones y conceptos similares).

Pueden efectuarse calificaciones, análisis y registros de calificaciones usando un lector óptico conectado a una microcomputadora que tenga los programas de evaluación apropiados. Sin embargo, los paquetes de cómputo que elaboran pruebas de acuerdo con ciertas especificaciones, las califican, analizan y presentan los resultados, son complejos y costosos. Un ejemplo de dichos programas para fines generales es MicroCAT (de Assessment Systems Corporation), el cual hace posible la construcción, aplicación, calificación y análisis de pruebas diseñadas a partir de la perspectiva de respuesta al ítem o de la evaluación clásica y administradas mediante procedimientos adaptativos o convencionales. MicroCAT crea y mantiene bancos de reactivos que consisten en texto, gráficas e imágenes digitalizadas; desarrolla y elabora formas de pruebas impresas; produce y aplica tests computarizados que van desde simples pruebas convencionales hasta complejas pruebas adaptativas, y realiza análisis de reactivos convencionales, análisis de respuesta a ítemes y calibraciones de reactivos. Algunas funciones de elaboración y administración de pruebas de MicroCAT están disponibles en línea y es posible acceder a ellas mediante programas de cómputo como los sistemas C-Quest y FastTEST proporcionados por Assessment Systems Corporation.

Errores humanos de calificación. La calificación de pruebas por computadora no es un proceso totalmente exento de errores, por ello se recomienda que los servicios de calificación de pruebas revisen la frecuencia de errores y emitan informes con las correcciones adecuadas cuando se encuentren tales fallas (American Educational Research Association *et al*., 1999). No obstante, en comparación con la calificación a mano, las tasas de errores de la calificación por computadora son reducidas.

Considerando el hecho de que las instrucciones para calificar muchas pruebas individuales de inteligencia y personalidad no siempre son claras y objetivas, no es sorprendente que lle-

guen a asignarse distintas puntuaciones a la misma respuesta. Aunque la variabilidad en las calificaciones tal vez sea mayor en el caso de evaluadores con poca experiencia (Slate y Jones, 1990), incluso los más experimentados cometen errores. Por ejemplo, se ha descubierto que los errores tanto en administración como en calificación ocurren cuando los estudiantes de psicología e incluso psicólogos profesionales administran pruebas de inteligencia individual (Franklin y Stillman, 1982; Ryan, Prefitera y Powers, 1983). En varios casos, los errores son de tal magnitud que se asignan a las personas niveles de inteligencia equivocados. También el personal clínico capacitado comete errores al calificar a mano inventarios de personalidad, en ocasiones tan graves que llegan a alterar los diagnósticos clínicos (Allard, Butler, Faust y Shea, 1995; Allard y Faust, 2000). Otros estudios han revelado que los resultados de las calificaciones se modifican por el agrado de quien administra o califica el examen hacia el examinado. También percibir al examinado como una persona cálida (Donahue y Sattler, 1971), brillante o aburrida (Sattler, Hillix y Neher, 1970; Sattler y Winget, 1970) puede afectar el resultado. Pueden ocurrir errores al convertir calificaciones brutas en calificaciones estándar o escaladas cuando se desconoce o se calcula mal la edad cronológica exacta del examinado.

Ponderación de calificaciones para reactivos de opción múltiple y de falso-verdadero. Parece razonable esperar que en pruebas objetivas, como en reactivos de ensayo, la cantidad de puntos asignada a una respuesta varíe de acuerdo con el tipo de reactivo y la calidad de la respuesta. Se han llevado a cabo muchos estudios sobre los efectos de la ponderación previa de las respuestas a reactivos de pruebas objetivas convencionales, es decir, asignar distinta cantidad de puntos a tipos de reactivos diferentes y a diversas respuestas. Algunas investigaciones han concluido que la ponderación previa es más definida y confiable que la calificación convencional (Hsu, Moss y Khampalikit, 1984; Serlin y Kaiser, 1978; Willson, 1982). Sin embargo, las ventajas de la ponderación diferencial de las respuestas a reactivos no parecen justificarse por el aumento en el costo y el tiempo de calificar (Kansup y Hakstian, 1975). En pruebas de 20 o más reactivos, asignar simplemente una calificación de 1 a cada respuesta correcta y 0 a las incorrectas resulta tan satisfactorio como usar pesos diferenciales. Así, las calificaciones posibles en una prueba de 50 reactivos de opción múltiple que haya sido calificada en forma convencional o de otra de falso-verdadero calificada mediante este procedimiento varían de 0 a 50.

Asignar pesos diferenciales a distintas respuestas puede ser más efectivo si el tipo de respuesta requerida fuese cambiado. Una variante interesante del formato de falso-verdadero es pedir a los examinados que indiquen qué tan seguros se sienten de sus respuestas. La tabla 3.1 ilustra

TABLA 3.1 Procedimiento de ponderación de confianza para reactivos de falso-verdadero

	El examinado señala que:	
La afirmación en realidad es:	**VERDADERA**	**FALSA**
La afirmación probablemente es verdadera	2	–2
La afirmación posiblemente es verdadera	1	0
No tengo idea	.5	.5
La afirmación posiblemente es falsa	0	1
La afirmación probablemente es falsa	–2	2

Fuente: Robert L. Ebel, *Measuring Educational Achievement*, © 1965, p. 131. Adaptado con permiso de Prentice Hall, Englewood Cliffs, NJ.

dicho procedimiento de *ponderación de la confianza* para reactivos de falso-verdadero. Aunque este procedimiento represente una mejora sobre la calificación convencional de 0-1 para reactivos de falso-verdadero, tal vez dicha calificación es satisfactoria para la mayoría de las pruebas de aula formadas por 30 o más reactivos.

Calificación de reactivos de clasificación. Así como con los reactivos de falso-verdadero y de opción múltiple, los reactivos de respuesta breve y de aparejamiento pueden calificarse asignando 1 punto a las respuestas correctas y 0 puntos a las erróneas y las omisiones. Debido a la gran cantidad de órdenes distintos en que puede colocarse un grupo de reactivos, la calificación de reactivos de clasificación presenta un problema especial. Por ejemplo, el error de asignar el segundo lugar a un reactivo al que de hecho corresponde el primero, no es tan grave como colocar dicho reactivo en cuarto lugar.

Las dos fórmulas que pueden usarse para calificar reactivos de clasificación son:

$$S_1 = c\left[1 - \frac{2\Sigma|d| }{c^2 - j}\right], \qquad \textbf{(3.1a)}$$

$$S_2 = c\left[1 - \frac{2\Sigma (d)^2}{c(c^2 - 1)}\right], \qquad \textbf{(3.1b)}$$

En estas fórmulas, c representa la cantidad de cosas clasificadas, las d son valores absolutos de las diferencias entre las posiciones asignadas por el examinado y las posiciones predeterminadas, y $j = 0$ cuando c es par y 1 si c es non. Para ejemplificar el uso de estas fórmulas, supongamos que deben ordenarse cinco ciudades de acuerdo con su población asignando una posición de 1 a la ciudad con la mayor población, 2 a la segunda más grande, y así sucesivamente. Los nombres de las cinco ciudades se incluyen en la primera columna de la tabla 3.2, las posiciones predeterminadas aparecen en la segunda columna y las asignadas por un examinado hipotético en la tercera columna. La cuarta columna contiene los valores absolutos de las diferencias entre la posición correcta para cada ciudad y las posiciones predeterminadas, y la quinta columna presenta el cuadrado de dichas diferencias. El total de los valores absolutos de las diferencias entre las posiciones del examinado y las posiciones predeterminadas es 10, y el total del cuadrado de las diferencias es 28. Sustituir $c = 5$, $\Sigma|d| = 10$, y $j = 1$ en la fórmula 3.1a produce $5[1 - 2(10)/(5^2 - 1)] = .83 \approx 1$. Sustituir $c = 5$ y $\Sigma d^2 = 28$ en la fórmula 3.1b da $5\{1 - 3(28)/[5(5^2 - 1)]\} = 1.5 \approx 2$. Los resultados de aplicar estas dos fórmulas no coinciden porque, comparada con la fórmula 3.1a, la fórmula 3.1b otorga más peso a las diferencias de posición mayores que a las menores. Cualquiera de las dos fórmulas es satisfactoria, dependiendo de si se opta por asignar un castigo extra a las respuestas que varían mucho de las predeterminadas. En cualquier caso, no hay un método único para calificar reactivos de pruebas que sea el mejor en todos los aspectos: eso depende de la filosofía y los objetivos del evaluador.

Corrección por adivinación. Después que la calificación total bruta se ha establecido, surge la pregunta de si es un indicador preciso de la verdadera situación del examinado en la prueba o si está inflada por los aciertos generados al adivinar. Es frecuente que las personas adivinen en pruebas objetivas, y las probabilidades de mejorar sus calificaciones de esa manera, en especial tratándose de reactivos con pocas opciones, pueden ser elevadas. Si la persona no conoce la respuesta correcta y todas las opciones son igualmente atractivas, la probabilidad de seleccionar la opción correcta adivinando es de $100/k$, donde k es el número de opciones por reactivo. Así, la posibilidad de adivinar la respuesta correcta es de 50 sobre 100 en un reactivo de falso-verdadero,

TABLA 3.2 Calificación de un ejemplo de reactivo de reacomodo

CIUDAD	RANGO CORRECTO	POSICIÓN DEL EXAMINADO	VALOR ABSOLUTO DE LA DIFERENCIA	CUADRADO DE LA DIFERENCIA
Houston	4	1	3	9
Chicago	3	2	1	1
Los Ángeles	2	3	1	1
Filadelfia	5	4	1	1
Nueva York	1	5	4	16
Totales			10	28

pero sólo de 25 sobre 100 en un reactivo de cuatro opciones. Obviamente, adivinar las respuestas de una gran cantidad de reactivos puede tener un efecto mucho más grave en una prueba de falso-verdadero que en una de opción múltiple.

Corregir los efectos de adivinación en ciertas pruebas estandarizadas (por ejemplo, la SAT y la GRE) conlleva restar una porción de la cantidad de respuestas erróneas a la cantidad de respuestas acertadas. El razonamiento en que se basan las fórmulas llamadas de corrección por adivinar no es de nuestro interés aquí, salvo en lo concerniente a la suposición cuestionable de que los examinados adivinan a ciegas cuando tienen dudas. La fórmula de corrección por adivinar más común es:

$$S = R - \frac{W}{k - 1}, \tag{3.2}$$

donde R es la cantidad de reactivos que el examinado acierta, W la cantidad de reactivos en que el examinado se equivoca, k la cantidad de opciones por reactivo y S la calificación corregida. Esta fórmula se ha criticado por producir resultados que son demasiado bajos cuando los examinados están menos familiarizados con el material de la prueba y muy elevados cuando están más familiarizados con dicho material (Little, 1962, 1966). Una fórmula alternativa propuesta por Little (1962) es:

$$S = R - \frac{W}{2(k - 1)}, \tag{3.3}$$

Los profesionales que administran pruebas están de acuerdo, por lo general, en que las fórmulas de corrección por adivinación en realidad no corrigen los efectos de adivinar y suelen tener poca influencia en el orden de las calificaciones. Hay excepciones cuando la cantidad de reactivos sin contestar varían mucho entre las personas y cuando algunos reactivos tienen más probabilidades de ser contestados que otros. Por lo regular, estas fórmulas, que suponen procedimientos similares a asignar pesos diferenciales a distintos reactivos, no se recomiendan para calificar pruebas de aula. Probablemente son más útiles para revisar pruebas de falso-verdadero y de velocidad, en las cuales el factor de adivinación interviene mucho más que en otro tipo de exámenes. Las calificaciones negativas, que en general se originan cuando se aplica la fórmula 3.2 en pruebas de falso-verdadero ($S = R - W$), usualmente se cambian por cero. De cualquier

modo, los examinados tienen derecho a saber si sus resultados se modificarán por adivinación. En las instrucciones de las pruebas debe añadirse información sobre cómo habrá de calificarse, incluyendo si se empleará corrección por adivinar.

Calificaciones modificadas. Usualmente no vale la pena alterar las calificaciones sin depurar de las pruebas objetivas mediante la ponderación diferencial de reactivos o con fórmulas de corrección por adivinar, pero a menudo se modifican de otras maneras para que resulten más significativas. Como se describe en la sección sobre normas del capítulo 4, el proceso de interpretar resultados de pruebas se facilita al transformarlos en calificaciones de percentiles o en calificaciones estándar.

Calificación de pruebas orales

Aunque es más probable que ocurran errores al calificar respuestas orales que escritas, hay formas especiales de evaluar el desempeño que mejoran la objetividad de la calificación en pruebas orales (vea la forma 3.1). Otras maneras de reducir los errores en este tipo de pruebas consisten en prestar atención al diseño de las preguntas, elaborar modelos de respuestas a las preguntas antes de administrar la prueba, recurrir a varios evaluadores y capacitar a los examinadores para evitar favoritismos y otros sesgos. Si el tiempo asignado a la calificación no es crítico, puede mejorarse su precisión si se graban las respuestas y vuelven a evaluarse más tarde (vea Aiken, 1983a).

FORMA 3.1 Forma para evaluar informes orales

Instrucciones: Para cada una de las preguntas de la lista, califique el desempeño del estudiante en una escala de 1 a 10: 1 corresponde a muy deficiente y 10 a excelente. Escriba el número adecuado (1 a 10) en la raya.

____ **1.** ¿Qué tan bien conoce el estudiante el tema del informe?
____ **2.** ¿Qué tan bien organizado estaba el informe?
____ **3.** ¿Qué tan eficaz fue la introducción para captar su atención?
____ **4.** ¿Con cuánta claridad y precisión habló el alumno?
____ **5.** ¿Qué tan interesante fue el tema?
____ **6.** ¿Qué tan eficaces fueron los materiales audiovisuales (películas, carteles, notas del pizarrón) en caso de haberse usado?
____ **7.** ¿Qué tanto se abstuvo el alumno de ver sus notas casi todo el tiempo y en cambio miró a la clase durante el informe?
____ **8.** ¿Con cuánta eficiencia usó el alumno gestos, posturas corporales y otros mensajes no verbales para comunicarse?
____ **9.** ¿En qué medida el estudiante se refirió a investigaciones u otras fuentes para presentar el informe?
____ **10.** ¿Cómo calificaría la conclusión (resumen de los puntos principales, preguntas para reflexionar, etc.) del informe?

Comentarios:

Calificación y notas

Después de haber administrado y calificado las pruebas, es preciso calificar los resultados. En el caso de pruebas realizadas por el maestro, la evaluación de los resultados en general implica asignar letras o notas. La asignación de notas es un proceso bastante subjetivo, dependiente no sólo de la prueba misma, sino de las expectativas del evaluador y de las calificaciones obtenidas por otros estudiantes. Algunos maestros califican estrictamente sobre la curva, mientras que otros evalúan en términos de un estándar o criterio de desempeño fijo. Sin embargo, la mayoría tal vez emplea una combinación de notas de curva y de estándar fijo. En un procedimiento de curva, el *método Cajori*, se asignan letras como sigue: A para el mejor 7% de las pruebas, B al siguiente 24%, C al 38% que sigue, D al 24% siguiente y F al 7% más bajo. La desventaja de este método es que no considera que la dificultad de las pruebas varía y que el nivel de capacidad promedio no es el mismo para estudiantes de distintas clases. Otro procedimiento de curva establece límites de notas con letra para pruebas de aula cuando el nivel de capacidad de la clase, el desempeño de la clase en la prueba con respecto a otras clases, y los propios resultados de la prueba se toman en cuenta (Aiken, 1983b, 2000).[3]

El sistema de asignación de notas, en que A se considera excelente o superior, B superior al promedio o bueno, C es el promedio, D es inferior al promedio o deficiente, y F es insuficiente o reprobado, es una forma de interpretación de resultados o de evaluación del desempeño. Todas las instituciones públicas y privadas tienen estándares que se espera cumplan sus alumnos, empleados o miembros. Los estándares pueden ser flexibles, pero en algún momento se evalúa el desempeño de los miembros de la organización. El castigo por obtener una evaluación negativa puede consistir en trabajo extra, degradación, suspensión o incluso expulsión. Las recompensas por una evaluación favorable incluyen premios, privilegios y ascensos.

Las notas en letra implican la evaluación del desempeño académico mediante la aplicación de diversas pruebas de aprovechamiento a los estudiantes. Las calificaciones en otras pruebas de habilidad y personalidad también requieren de interpretación si se pretende usarlas para ciertos fines como ubicación en puestos o clases especiales, psicodiagnóstico o tratamientos psicológicos, u otro tipo de intervenciones. Interpretar las calificaciones en dichas pruebas puede ser un proceso muy complejo, dependiendo del tipo de prueba y los propósitos para los que se aplique. La interpretación involucra factores tanto objetivos como subjetivos, incluyendo el uso de normas como se analiza en el siguiente capítulo.

RESUMEN

Los procedimientos para administrar y calificar pruebas varían en cierta medida según el tipo de prueba y las personas a las que está dirigida. Los examinados deben estar preparados, motivados para desempeñarse bien y relativamente exentos de tensión y de otras condiciones distractoras.

[3]Aiken (2000) analiza los problemas relacionados con las notas en letra y describe un conjunto de siete programas de cómputo que proporcionan una base más objetiva para la asignación de notas. Estos programas pueden usarse para asignar letras, calcular el promedio o una serie de letras ponderadas, convertir letras en percentiles, transformar notas en puntos de calidad y calcular estadísticas apropiadas, convertir calificaciones numéricas en una escala diferente, puntos en percentiles y calificaciones estándar, y almacenar o recuperar notas en letras, números enteros o decimales de un archivo. Previa solicitud puede obtenerse una copia de esta serie de programas enviando un disquete formateado de sistema DOS y un sobre de porte pagado al doctor Lewis R. Aiken, 3300 Blue Ridge Court, Thousand Oaks, CA 91362.

Quienes administran las pruebas deben estar capacitados, familiarizados con la prueba en particular y tener la seguridad de que todo está en orden antes de iniciar una prueba. En general, las circunstancias de prueba deben ser física y psicológicamente cómodas, de modo que los examinados se sientan dispuestos a realizar su mejor esfuerzo.

Como regla general, debe informarse a los examinados sobre el o los objetivos de la prueba, cuándo y dónde se administrará,cuál será el formato y el material que aborda. Los examinadores deben seguir las instrucciones cuidadosamente, tomar precauciones para reducir al mínimo las trampas y prepararse para manejar emergencias y otros problemas especiales. Suele permitirse cierta flexibilidad al aplicar pruebas elaboradas por maestros y estandarizadas, pero en caso de alejarse radicalmente de las instrucciones de administración se invalida el uso de las normas en las pruebas estandarizadas. Los examinadores también deben intentar entablar un buen *rapport* con los examinados, en particular en el caso de pruebas aplicadas de manera individual.

El ingenio para resolver pruebas, los aciertos al adivinar, cambiar las respuestas y hacer trampas son algunos de los factores que pueden inflar los resultados en una prueba objetiva; alardear, usar una redacción rebuscada o buena caligrafía tienen el mismo efecto en las pruebas de ensayo. La influencia del ingenio lpara resolver pruebas se minimiza al elaborar los reactivos con cuidado y evitando las claves como la extensión de los reactivos, determinantes específicos, errores gramaticales, indicios estilísticos y opciones heterogéneas (no paralelas). Con frecuencia se aplican fórmulas de corrección por adivinación para reducir los efectos de adivinación. No obstante, con la posible excepción de los reactivos de falso-verdadero, al calificar pruebas de aula, las pruebas convencionales de corrección por adivinación no suelen compensar el tiempo y los esfuerzos invertidos.

Las pruebas de ensayo pueden calificarse holística o analíticamente, pero en ambos casos debe informarse a los examinados cómo se revisarán las pruebas. Se recomienda calificar las respuestas de todos los examinados a una pregunta específica antes de continuar con la siguiente, lo mismo que evaluar el contenido y el estilo de las respuestas en forma separada. Además de una calificación numérica, a menudo es útil incluir comentarios, correcciones y explicaciones como retroalimentación sobre el desempeño en las pruebas de ensayo.

Muchas pruebas objetivas se califican con ayuda de computadoras u otras máquinas especiales. En general, la calificación a máquina es superior en términos de velocidad y precisión, pero menos flexible que la realizada a mano. La evaluación de muchos tests de inteligencia y personalidad individuales no es del todo objetiva, y pueden cometer errores graves tanto los profesionales como el personal capacitado.

En las pruebas adaptativas, donde la secuencia de las preguntas presentadas al examinado varía de acuerdo con su posición estimada en la variable especificada y con sus respuestas a reactivos anteriores, el tiempo de administración se reduce considerablemente. El uso de computadoras para presentar reactivos y evaluar respuestas hace de las pruebas adaptativas una opción eficiente, aunque más costosa, que el método tradicional de presentar dichos reactivos a todos los examinados.

Se han investigado a fondo los efectos de asignar distinto peso en la calificación para diferentes tipos de reactivos objetivos o para distintas respuestas a un reactivo. En general, no se recomiendan ponderaciones previas para calificar pruebas que consistan en 20 o más reactivos.

Las calificaciones brutas con frecuencia se convierten en percentiles o calificaciones estándar con el fin de calcular porcentajes, realizar comparaciones e interpretar calificaciones. Las calificaciones en pruebas de aula también pueden convertirse en notas, ya sea usando un conjunto establecido de porcentajes como los especificados en el método Cajori o de una manera más subjetiva.

PREGUNTAS Y ACTIVIDADES

1. Defina lo que es *el ingenio para resolver pruebas* y describa los comportamientos que revelan dicha conducta. ¿Qué puede hacer un diseñador de pruebas para reducir lo más posible los efectos de tal habilidad en las calificaciones?

2. Pregunte a un grupo de sus compañeros de clase sobre las técnicas que usan para elegir respuestas en pruebas con reactivos de opción múltiple cuando no han estudiado el material en forma adecuada. ¿Qué técnicas son más comunes y qué tan efectivas son?

3. Sin duda ha observado que la velocidad para resolver una prueba de aula puede variar en gran medida de estudiante a estudiante. Algunos terminan un examen de dos horas en menos de una hora, mientras que otros continúan trabajando después de terminado el tiempo permitido. A juzgar por sus observaciones y conversaciones, ¿cuáles considera que son los principales factores que determinan la velocidad para concluir una prueba?

4. ¿Qué es una prueba adaptativa? ¿De qué manera las pruebas adaptativas son mejores que los procedimientos de evaluación objetivos convencionales? ¿En qué aspectos son inferiores?

5. ¿Cuáles son algunas de las ventajas y desventajas de elaborar, administrar y calificar pruebas por computadora, en comparación con las mismas actividades realizadas mediante procedimientos convencionales?

6. Juan resuelve una prueba de 50 reactivos de opción múltiple, con cuatro opciones. Acierta en 30 reactivos, se equivoca en 16 y deja 4 sin contestar. ¿Cuál es su calificación total, con corrección por adivinación y sin ésta? Si todos los reactivos son de falso-verdadero y obtiene el mismo número de aciertos y errores ya mencionados, ¿cuál será su calificación total, con y sin corrección por adivinación?

7. Un examen sobre historia británica contiene un reactivo de reacomodo consistente en una lista de siete batallas. Se pide a los alumnos que ordenen las siete batallas de acuerdo con la fecha en que ocurrieron. El orden correcto es: Batalla de Hastings, Batalla de Bunker Hill, Batalla de Yorktown, Batalla de Trafalgar, Batalla de Waterloo, Batalla del Marne, Battalla de Bretaña. Juan ordena las batallas de la siguiente manera: Waterloo, Hastings, Yorktown, Trafalgar, Marne, Bretaña y Bunker Hill. ¿Cuál sería su calificación en este reactivo? María elige el siguiente orden: Hastings, Waterloo, Yorktown, Bunker Hill, Trafalgar, Marne, Bretaña. ¿Cuál es su calificación?

8. Usando los porcentajes diseñados mediante el método Cajori, asigne letras de grado a las calificaciones de la distribución *X* en el ejercicio 3 del apéndice A (página 446). Después asigne letras de grado a la distribución *Y* del mismo ejercicio. Suponga que la calificación máxima es 50, la mínima 0, y la habilidad mediana de la clase es 50.

9. Observe la administración de una prueba en una de sus clases. ¿El examinador siguió los lineamientos descritos en este capítulo? Si no fue así, ¿qué errores cometió y cuáles fueron las consecuencias reales o posibles de sus equivocaciones?

CAPÍTULO CUATRO

ANÁLISIS DE REACTIVOS Y ESTANDARIZACIÓN DE PRUEBAS

Este capítulo aborda dos temas en cierta medida técnicos, pero importantes: el análisis de reactivos y la estandarización de pruebas. Ambos temas tienen que ver con el cálculo de ciertos análisis estadísticos que deben revisarse con detalle para determinar si todos los reactivos de una prueba están funcionando como deberían, y cómo pueden interpretarse las calificaciones de las pruebas. El análisis de reactivos se centra en el funcionamiento de reactivos individuales, mientras que la estandarización de pruebas se ocupa de la interpretación normativa de los resultados de la prueba en su conjunto o de algunas de las partes o subpruebas que la integran. Los temas de este capítulo y del siguiente se tratan, sobre todo, desde la perspectiva de la teoría clásica (tradicional) sobre pruebas, pero no se dejan de lado los enfoques más recientes de la teoría de respuesta al ítem. Tanto la teoría clásica sobre pruebas (CTT) como la teoría de la respuesta a los ítemes (IRT) son útiles para el desarrollo, el análisis y las aplicaciones de pruebas y, dependiendo de la tarea específica, ambas han recibido apoyo.

ANÁLISIS DE REACTIVOS

Incluso después de haber sido administrada y calificada una prueba, no siempre es seguro que haya funcionado bien. Cuando se pilotea una prueba en un principio, es posible que surjan varios problemas. Ésta es una de las razones de que las pruebas que se distribuyen comercialmente se administren primero a una muestra de personas representativas del grupo que las pruebas están destinadas a medir. Entonces pueden analizarse las respuestas de esa muestra piloto para determinar si los reactivos están funcionando de manera adecuada.

Cualquiera que sea el tipo de prueba, estandarizada o elaborada por el maestro, de habilidad o de personalidad, un análisis *post mortem* o *post hoc* de los resultados es tan necesario como en medicina o en cualquier otra empresa humana. Entre las preguntas que es preciso contestar figuran las siguientes: ¿fueron adecuados los límites de tiempo? ¿Los examinados entendieron las instrucciones? ¿Fueron apropiadas las condiciones en que se administró la prueba? ¿Se manejaron de manera adecuada las situaciones de emergencia? Es inusual que puedan anticiparse todos los problemas o contingencias que surgen durante un piloteo, pero un análisis posterior puede proporcionar información y motivación para prever y manejar situaciones similares al administrar pruebas en el futuro. El cuestionario de la forma 4.1, que responden los examinados inmediatamente después de haberse sometido a una prueba de aprovechamiento, puede ofrecer información cualitativa sobre las percepciones en cuanto a la imparcialidad de la prueba, si se sentían preparados para ella, si cumplió con sus expectativas o cómo respondieron a los reactivos individuales.

FORMA 4.1 Forma de evaluación de test

Instrucciones: Llene esta forma después de terminar la prueba. Encierre su respuesta en un círculo para cada reactivo y responda en los espacios en blanco de ser necesario.

Sí No 1. ¿Fue satisfactorio el ambiente (asientos, temperatura, ventilación, nivel de ruido, etc.) en que se aplicó la prueba? _____

_____

Sí No 2. ¿Leyó usted cuidadosamente las instrucciones antes de empezar la prueba?

Sí No 3. ¿Fueron claras las instrucciones?

Sí No 4. ¿El formato de la prueba (tipo de reactivos, acomodamiento, hoja de respuestas) fue satisfactorio? _____

_____

Sí No 5. ¿La prueba cubrió de manera adecuada el material asignado? _____

_____

Sí No 6. ¿Las preguntas de la prueba tenían dificultad adecuada? _____

_____

Sí No 7. ¿Estudió usted lo suficiente para la prueba? _____

_____

Sí No 8. ¿Estudió el material correcto? _____

_____

Sí No 9. ¿Piensa que respondió las preguntas de manera equivocada? ¿Cuáles? ___

_____

Sí No 10. ¿Adivinó algunas de las respuestas? ¿Cuántas? ¿Cuáles? _____

_____

Sí No 11. ¿Omitió usted algunos de los reactivos? ¿Cuáles? _____

Sí No 12. ¿Tuvo bastante tiempo para terminar la prueba? _____

_____

Sí No 13. ¿Al terminar la prueba, revisó sus respuestas?

Sí No 14. ¿Estuvo nervioso o emocionalmente molesto durante la prueba?

Sí No 15. ¿Fue justa la prueba? _____

_____

Sí No 16. En general, ¿considera que la prueba fue buena? _____

_____

Sí No 17. ¿Durante la prueba observó que se hiciera trampa? _____

_____

18. ¿Qué calificación espera obtener en esta prueba? _____

El análisis de las respuestas que da un grupo determinado de personas a un reactivo individual en una prueba cumple varias funciones. El principal objetivo de dicho *análisis de reactivos* es contribuir a mejorar la prueba al revisar y descartar reactivos ineficaces. Otra función importante de dicho análisis, en especial en una prueba de aprovechamiento, es proporcionar información diagnóstica sobre lo que saben o no los examinados.

Pruebas con referencias a criterios y de dominio

El procedimiento empleado en evaluar la eficacia de los reactivos de prueba depende, en cierta medida, del propósito de la misma. Por ejemplo, el examinador puede estar interesado sólo en determinar qué tanto sabe un examinado sobre el material de la prueba, no en comparar su desempeño con el de otras personas. En este caso, el desempeño se mide contra un criterio o estándar establecido por el maestro del aula o por una política institucional. El objetivo de tal evaluación con *referencias a criterio* (o a un área) no es descubrir qué calificación obtiene una persona en relación con otras, sino en qué nivel se encuentra en cuanto a determinados objetivos de una lección, curso o programa. Un tipo particular de prueba con referencias a criterio, diseñada para medir el logro de un rango limitado de habilidades cognoscitivas, se conoce como *prueba de dominio*. La calificación de una persona en una prueba de dominio, o en cualquier otra prueba con referencias a criterio, se expresa como un porcentaje de la cantidad total de reactivos respondidos de manera correcta; una calificación perfecta indica el 100% de dominio del material.

Diferencias individuales y validez de los reactivos

Dado que suele ser difícil llegar a un acuerdo sobre cuánto debe saber una persona sobre una materia en particular o en qué consiste dominarla, tradicionalmente las calificaciones se han interpretado comparándolas con las obtenidas por otras personas. Las pruebas psicológicas se han diseñado, sobre todo, para evaluar diferencias entre individuos en cuanto a características. Las habilidades y la personalidad de la gente difieren, y los psicólogos intentan evaluar estas diferencias mediante diversos tipos de pruebas. Mientras mayor sea el cuidado con que se lleva a cabo dicha evaluación, mayor será la precisión con que podrá predecirse el comportamiento a partir de los resultados de las pruebas. En consecuencia, los encargados de elaborar pruebas intentan diseñar reactivos que permitan diferenciar a las personas en cuanto a lo que se quiera medir. De esta manera, aumenta la variabilidad de los resultados totales de las pruebas y entonces una calificación determinada se convierte en un índice más preciso de la posición de una persona en relación con la de otros individuos.

A fin de evaluar la utilidad de un reactivo como medida de las diferencias individuales en cuanto a las características de habilidad o de personalidad, se requiere un criterio externo de medida de dicho rasgo. Si la prueba se elabora para predecir el desempeño en un trabajo o en la escuela, entonces un criterio apropiado consiste en la medida del desempeño laboral (digamos, las escalas del jefe) o del aprovechamiento escolar (por ejemplo, notas asignadas por el maestro). La *validez* de un reactivo para predecir una posición con base en un criterio externo puede determinarse al correlacionar las calificaciones de un reactivo (0 para los errores y 1 para los aciertos) con las calificaciones de la medida de criterio. Se han usado distintos tipos de coeficientes de correlación para este propósito; el más común es el *coeficiente biserial puntual*, que puede calcularse con la siguiente fórmula:

$$r_{pb} = \frac{(Y_p - Y)\sqrt{n_t n_p/[(n_t - n_p)(n_p - 1)]}}{S_t}, \qquad \textbf{(4.1)}$$

donde n_t = la cantidad total de examinados, n_p = la cantidad de examinados que resuelven correctamente el reactivo, = la media de las calificaciones de criterio de quienes pasan el reactivo, = la media de todas las calificaciones de criterio, y s_t = la desviación estándar de todas las calificaciones de criterio. El criterio puede ser externo (productividad en el trabajo o grados de un curso) o incluso calificaciones totales de la propia prueba.

Para ilustrar el cálculo del coeficiente biserial puntual, supongamos que la media y la desviación estándar del total de las calificaciones de un grupo de 30 personas son 75 y 10, respectivamente. Ahora bien, si la calificación media de 17 examinados que aciertan en determinado reactivo es 80, la sustitución de estos valores en la fórmula 4.1 produce:

$$r_{pb} = \frac{(80 - 75)\sqrt{30(17)/[13(29)]}}{10} = .58.$$

Cuanto más elevada sea la correlación entre reactivo y criterio, más preciso será el reactivo como predictor del criterio. El que un reactivo se conserve o deseche depende del tamaño de este coeficiente. Aunque reactivos con coeficientes tan bajos como .20 pueden contribuir a predecir el criterio, se prefieren coeficientes más elevados. Un reactivo con una correlación cercana o menor que .00 con el criterio debe, sin duda, revisarse o descartarse. Sin embargo, la utilidad de un reactivo para predecir un criterio específico no sólo depende de la correlación entre reactivo y criterio, sino también de la correlación del reactivo con otros reactivos de la prueba. Son mejores los reactivos que tienen correlaciones elevadas con el criterio, pero bajas con otros reactivos, porque representan una contribución más independiente a la predicción de calificaciones de criterio.

Dificultad de los reactivos e índices de discriminación

Por lo general, no hay un criterio externo fácilmente disponible contra el cual validar los reactivos de las pruebas de aprovechamiento en el aula, de modo que a menudo se emplea un procedimiento distinto, el de *consistencia interna*. Al igual que con cualquier otra prueba, el análisis de reactivos de una prueba de aula conlleva determinar el porcentaje de examinados que pasan el reactivo y la correlación del reactivo con una medida de criterio. No obstante, en el caso de una clase de aprovechamiento de aula, el criterio consiste en calificaciones totales sobre la prueba misma. Suponiendo que la serie de reactivos en conjunto es una medida adecuada de aprovechamiento en el sujeto, la suma de las calificaciones se usa como el criterio para determinar la consistencia interna de la prueba.

Un procedimiento más breve consiste en dividir a los examinandos en tres grupos según sus calificaciones en la prueba como un todo: un grupo superior formado por el 27% que obtuvo las calificaciones más altas, un grupo inferior compuesto por el 27% que tuvo las calificaciones más bajas, y el restante 46% incluido en un grupo intermedio. Cuando el número total de examinados es pequeño, el 50% correspondiente a los grupos inferior y superior a menudo se utiliza para propósitos de análisis de reactivos. En cualquier caso, los siguientes índices estadísticos se calculan a partir de los resultados de los grupos inferior y superior:

$$p = \frac{U_p + L_p}{U + L} \tag{4.2}$$

y

$$D = \frac{U_p - L_p}{U}. \tag{4.3}$$

U_p y L_p son la cantidad de individuos que hay en los grupos superior e inferior, respectivamente, y que aciertan en el reactivo; U y L son el número total de personas en los grupos superior e inferior (obsérvese que $U = L$), respectivamente. Al valor de p se le conoce como *índice de dificultad del reactivo* y al de D como *índice de discriminación del reactivo*. Para ejemplificar el cálculo de estos índices, supongamos que 50 personas presentan una prueba. Entonces los gru-

pos superior e inferior pueden formarse con los $.27 \times 50 \approx 14$ superior y el 14 inferior de la suma total de calificaciones. Si 12 de las personas del grupo superior y 7 de las que forman el grupo inferior pasan el reactivo A, entonces $p = (12 + 7)/28 = .68$ y $D = (12 - 7)/14 = .36$.

El índice de dificultad del reactivo tiene un rango de .00 a 1.00. Un reactivo con $p = .00$ es uno que nadie contestó correctamente, y un reactivo de $p = 1.00$ es el que todos respondieron en forma acertada. El valor *p* óptimo para un reactivo depende de varios factores, incluyendo los objetivos de la prueba y la cantidad de opciones de respuesta. Si el propósito de una prueba es identificar o seleccionar sólo un pequeño porcentaje de los mejores candidatos, entonces la prueba debe ser bastante difícil, como se refleja en un valor promedio inferior de *p*. Si la prueba está diseñada para rechazar sólo a algunos candidatos muy deficientes, entonces es mejor un valor promedio de *p* elevado. Por ejemplo, el valor de *p* óptimo debe ser muy bajo para reactivos de una prueba diseñada para otorgar becas o hacer contrataciones en puestos superiores, pero muy alto en una prueba diseñada para identificar estudiantes candidatos a programas terapéuticos. En una prueba elaborada para medir un rango amplio de habilidad, el valor de *p* óptimo es más cercano a .50. Como se muestra en la tabla 4.1, para una prueba semejante el valor promedio de *p* óptimo también varía inversamente con el número de opciones de respuestas (*k*). Los valores de *p* para reactivos aceptables caen dentro de un rango bastante estrecho, aproximadamente de .20, alrededor de estos valores registrados.[1] Aunque algunos reactivos muy fáciles y otros muy difíciles con frecuencia se incluyen en una prueba de rango amplio, de hecho agregan muy poco a la efectividad general para distinguir entre estudiantes que poseen distinta cantidad de conocimiento, habilidad o comprensión del material de prueba.

El índice de discriminación del reactivo (*D*) es una medida de la eficacia de un reactivo para discriminar entre quienes obtienen altas y bajas calificaciones en una prueba. Mientras más elevado sea el valor de *D*, resulta más eficaz para establecer dicha distinción. Cuando (*D*) es igual a 1.00, todos los examinados del grupo superior y ninguno del grupo inferior en las calificaciones totales de la prueba respondieron el reactivo en forma adecuada. Sin embargo, casi nunca resulta *D* igual a 1.00 y, por lo regular, se considera aceptable un reactivo si tiene un índice *D* de .30 o mayor. Pero *D* y *p* no son índices independientes, y el valor de *D* mínimo acepta-

TABLA 4.1 Índices medios óptimos de la dificultad de los reactivos para pruebas con reactivos de opción múltiple

NÚMERO DE OPCIONES (*k*)	ÍNDICE MEDIO ÓPTIMO DE DIFICULTAD (*p*)
2	.85
3	.77
4	.74
5	.69
Abierta (ensayo, respuesta breve)	.50

Fuente: Elaborado con datos proporcionados por F. M. Lord, *Psychometrika*, *17* (1952), pp. 181-194.

[1]El rango de *p* debe ser menor que .20 en una prueba con *topes máximos* diseñada para medir con eficacia dentro de un rango bastante estrecho de capacidad. Éste es el caso, por ejemplo, de una prueba diseñada para seleccionar o identificar un grupo de personas relativamente pequeño con habilidades muy bajas o muy altas o con cualesquier características que tengan una tasa de aparición baja (tasa base) en la población de interés.

ble varía de acuerdo con el valor de *p*. Un valor de *D* en cierta medida inferior a .30 es aceptable mientras *p* cada vez aumenta o disminuye más que el valor óptimo, sobre todo cuando los grupos de comparación superior e inferior son numerosos. Asimismo, un reactivo con un índice *D* bajo no se descarta automáticamente: es posible salvarlo modificándolo. Elaborar reactivos de pruebas adecuados es un proceso minucioso, de modo que los defectuosos deben corregirse y conservarse siempre que sea posible.

Factores que afectan el funcionamiento de los reactivos

Los resultados de un análisis de reactivos a menudo varían considerablemente dependiendo del grupo específico que se somete a la prueba, en particular cuando la cantidad de examinados es reducida. Algunos reactivos pueden responderse de manera diferente por hombres y por mujeres o por algún grupo étnico, de edad o socioeconómico en comparación con otro. Al elaborar una prueba estandarizada, en la actualidad es frecuente revisar cada reactivo y los análisis estadísticos correspondientes para buscar indicios de falta de discriminación o sesgo por grupo. Para facilitar este proceso, a menudo se calculan índices estadísticos del *funcionamiento diferencial del reactivo (DIF).* Se han propuesto muchos métodos para obtener información sobre el funcionamiento diferencial de los reactivos de pruebas, incluyendo el procedimiento de diagramas delta de reactivos del Servicio de Evaluación Pedagógica y varios procedimientos de chi cuadrada (vea Camilli y Shepard, 1994).

Sólo porque la forma como se responde un reactivo varía de grupo a grupo no quiere decir necesariamente que un reactivo esté sesgado en contra de alguno de los grupos. Técnicamente, un reactivo se considera sesgado sólo cuando mide algo distinto —una característica o rasgo diferente— en un grupo con respecto a otro. Si las calificaciones de un grupo reflejan diferencias verdaderas en cuanto a la capacidad o cualquier característica para cuya medición se diseñó el reactivo, éste se encuentra técnicamente libre de sesgo. Al realizar un análisis de reactivo individual para cada grupo puede revelarse la presencia de sesgo en el reactivo, es decir, si el reactivo discrimina bien entre calificaciones altas y bajas en ambos grupos.

También surgen problemas en el análisis de reactivos de las pruebas de velocidad, en las que los límites de tiempo son breves y no todos los examinados pueden terminar. En una prueba de velocidad, los reactivos cercanos al final de la prueba intentan resolverse por relativamente pocas personas. Si quienes alcanzan y por consiguiente tratan de resolver un reactivo final son los examinados más capaces, el índice de discriminación (*D*) probablemente será mayor del que resultaría si el límite de tiempo fuera más generoso. Por otra parte, si los más descuidados tienen más probabilidades de llegar a los reactivos del final de la prueba e intentar responderlos, los valores *D* de dichos reactivos tenderán a ser inferiores a los de aquellos que se encuentran cercanos al principio. Se han propuesto varios procedimientos para resolver los problemas que genera el análisis de reactivos hacia el final de las pruebas de velocidad, pero ninguno resulta del todo satisfactorio.

A pesar de sus desventajas, los índices de dificultad y de discriminación de reactivos proporcionan información útil sobre el funcionamiento de los reactivos individuales. En general, se ha descubierto que el análisis de reactivos produce mejoras considerables en la eficacia de las pruebas. En particular, el índice de discriminación de reactivos es una medida bastante adecuada de la calidad del reactivo. Junto con el índice de dificultad (*p*), *D* puede servir como una advertencia de que algo está fallando en un reactivo.

Los constructores de pruebas a menudo han recibido el consejo de registrar los resultados estadísticos del análisis de reactivos, junto con el reactivo mismo, en tarjetas de índices y archivar las tarjetas para su uso posterior. Con la llegada de las computadoras de alta velocidad, ahora los reactivos pueden codificarse por tema, niveles de dificultad y de discriminación, y tal vez

hasta por los procesos cognoscitivos que implica responderlos, y después almacenarlos en un banco de reactivos. No sólo los profesionales que elaboran pruebas usan estos bancos de reactivos, también están disponibles como complementos de muchos libros de texto para usarse como pruebas prácticas o servir como banco de reactivos al elaborar pruebas de aula. Las computadoras pueden utilizarse para seleccionar reactivos de un banco que maneje distintos contenidos e integrarlos como una unidad de prueba o exámenes. También hay programas de computación especializados para facilitar la elección de reactivos que abordan un tema específico y, además, con las características estadísticas deseadas.

Consistencia interna contra validez

El concepto de validez del reactivo, en general, se refiere a la relación entre un reactivo y un criterio externo. Pero *D* es una medida de la relación de los resultados de reactivos con un criterio interno (total de calificaciones de la prueba) más que con un criterio externo. Seleccionar reactivos con valores *D* altos dará como resultado una prueba internamente consistente en la que las correlaciones entre reactivos son muy positivas. Sin embargo, las calificaciones de una prueba internamente consistente no siempre están muy correlacionadas con las calificaciones de un criterio externo. Para construir una prueba con una elevada correlación con un criterio externo, deben seleccionarse reactivos que tengan correlaciones bajas entre sí, pero elevadas con la medida de criterio. Seleccionar reactivos con base en la estadística *D* origina un tipo de prueba distinto al de una prueba compuesta por reactivos elegidos por sus altas correlaciones con un criterio externo. Cuál de estas estrategias, interna o externa, es superior depende de los propósitos de la prueba. Si se desea una medida internamente consistente de una característica, debe usarse el índice de discriminación (*D*) para seleccionar reactivos. Si se requiere el predictor más válido de un criterio externo en particular, deberán emplearse las correlaciones de criterio de reactivos. En ocasiones es adecuada una combinación de ambas estrategias: se elabora una prueba compuesta a partir de subpruebas con bajas correlaciones entre sí y correlaciones considerables con un criterio externo, pero los reactivos de cada subprueba están altamente intercorrelacionados.

Reactivos con referencias a criterios

Los índices de dificultad y de discriminación pueden calcularse también en reactivos de prueba con referencia a criterios, y se diseñan para determinar las posiciones de los examinados en objetivos pedagógicos específicos. En este caso, los examinados se dividen en dos grupos: un grupo superior consistente en los examinados *U*, cuyas calificaciones totales de prueba cumplen con el criterio establecido de desempeño aceptable, y un grupo inferior integrado por los examinados *L*, cuyas calificaciones totales no satisfacen los criterios. Para un reactivo particular, U_p es el número en el grupo superior (encima del nivel de criterio) de quienes aciertan en el reactivo, y L_p es el número en el grupo inferior (debajo del nivel de criterio) de los que aciertan en el reactivo. Entonces el índice de dificultad del reactivo se define mediante la fórmula 4.2. Debido a que *U* y *L* no necesariamente son iguales, el índice de discriminación del reactivo se define como:

$$D = \frac{U_p}{U} - \frac{L_p}{L}. \qquad \textbf{(4.4)}$$

Puede emplearse un criterio externo para formar los grupos superior e inferior. En el caso de una prueba de aprovechamiento con referencia a criterio, por ejemplo, los examinados pueden dividirse en dos grupos: los que recibieron instrucciones sobre el tema asociado con la prueba (*U*) y quienes no recibieron dichas instrucciones (*L*). Los grupos *U* y *L* también pueden consis-

tir en los mismos individuos, tanto antes (*L*) como después (*U*) de la instrucción. En cualquier caso, puede usarse la fórmula 4.4 para determinar un índice de discriminación de reactivos.

Análisis de distractores

El análisis de los reactivos de opción múltiple suele empezar con el cálculo de índices de discriminación y dificultad para cada reactivo. Un análisis secundario se ocupa del funcionamiento de los distractores $k - 1$ para cada reactivo. El índice de discriminación de reactivos (*D*) proporciona cierta información sobre el funcionamiento de los distractores en conjunto. Un *D* positivo indica que los examinados en el grupo superior (en la calificación total de la prueba) tendieron a seleccionar uno de los distractores; la magnitud de *D* indica la medida de esta tendencia. Por otra parte, un *D* negativo indica que los distractores se eligieron con mayor frecuencia por examinados del grupo superior que por los del grupo inferior y que el reactivo debe revisarse. Sin embargo, el signo y la magnitud de *D* no revelan si todos los distractores funcionaron de manera adecuada.

El método más sencillo para determinar si todos los distractores están funcionando como deberían es contar el número de veces que cada distractor se seleccionó como la respuesta adecuada por los examinados del grupo superior y por los del grupo inferior. Si, en el caso de un reactivo que por lo demás es satisfactorio, demasiados examinados del grupo superior o muy pocos del grupo inferior seleccionaron un distractor determinado, éste debería ser modificado o reemplazado. En términos ideales, todos los distractores $k - 1$ deberían ser igualmente aceptables para los examinados que no conocen la respuesta correcta de un reactivo; en consecuencia, todo distractor debe ser seleccionado por alrededor de la misma cantidad de personas.

Curvas características de los reactivos

Incluso los valores aceptables de *p* y *D* no garantizan que un reactivo esté funcionando de manera efectiva a lo largo de todos los niveles de desempeño de la prueba. Para ser más efectivo, la proporción de las personas que contestan un reactivo correctamente debería aumentar en forma continua con el incremento de las calificaciones totales en la prueba o subprueba. El que un reactivo de prueba funcione de esta manera puede determinarse mediante la *curva característica del reactivo (ICC).* Al construir una ICC, la proporción de examinados que dieron la respuesta en clave se contrasta contra sus calificaciones en un criterio interno (por ejemplo, las calificaciones totales de la prueba) o un criterio externo, como el aprovechamiento académico o el desempeño laboral. Una vez que se ha construido la curva característica de un reactivo en particular, es posible determinar el nivel de dificultad y el índice de discriminación de dicho reactivo. El *nivel de dificultad (b)* es la calificación de criterio en el que 50% de los examinados dio la respuesta acertada (predeterminada); el *índice de discriminación (a)* es la pendiente de la curva característica del reactivo en el punto del 50%. Por ejemplo, de las dos curvas características del reactivo trazadas en la figura 4.1, un valor de .50 en el eje vertical corresponde a una calificación total en la prueba de 68 en el caso del reactivo 1 y de 77 en el reactivo 2. Por consiguiente, el reactivo 2 es más difícil que el 1. Sin embargo, la ICC del reactivo 1 tiene una pendiente más pronunciada que la del reactivo 2, de modo que el reactivo 1 discrimina mejor que el 2 entre quienes obtienen las calificaciones superiores y los de las calificaciones inferiores en toda la prueba. Estas dos medidas (ubicación y pendiente de la ICC) son similares a los índices *p* y *D* del análisis de reactivos tradicional, pero una ICC proporciona de mejor modo una imagen detallada del funcionamiento de reactivos a lo largo de todo el rango de calificaciones de criterio interno o externo. Además

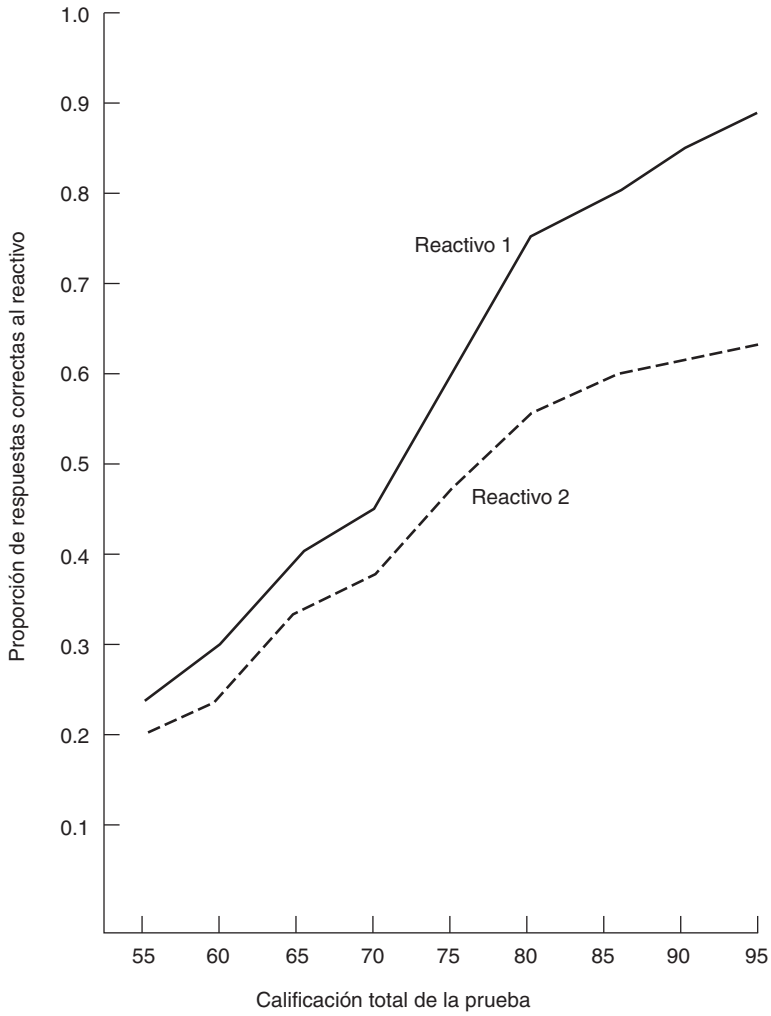


FIGURA 4.1 Dos curvas características de reactivos.
Vea la explicación en el texto.

de trazar la proporción de respuestas correctas que corresponden a las calificaciones totales de una medida de criterio externo o interno, la proporción de individuos que obtiene cada calificación y que seleccionaron un distractor en particular puede trazarse al analizar la eficacia de los distractores de reactivos.

Teoría de respuesta al Ítem

A diferencia de la atención más bien superficial que otorgan la teoría y los métodos tradicionales sobre pruebas a las respuestas a reactivos individuales, tales respuestas son el núcleo de la teoría y metodología de respuesta a los ítemes. La *teoría de respuesta al Ítem* (*IRT*) se basa en

la relación funcional teórica entre un continuo de capacidad latente supuesto y las respuestas a reactivos individuales en una prueba. Los cálculos que conlleva son muy complicados y en general deben realizarse con la ayuda de un programa de cómputo como LOGIST, BILOG, ASCAL o BIGSTEPS (vea Mislevy y Stocking, 1989; Vale, 1985; Wright y Linacre, 1991).

El modelo usual de la IRT es una función logística que tiene uno, dos o tres parámetros. La fórmula para generar cálulos de probabilidad en el *modelo de tres parámetros* es:

$$P(\theta) = c + (1 - c)\frac{1}{1 + e^{-a(\theta - b)}}. \quad \textbf{(4.5)}$$

En esta fórmula, e es la base del logaritmo natural (2.718282), a es un parámetro de la pendiente de un reactivo, b es un parámetro de ubicación de un reactivo, c es un parámetro seudoadivinatorio, θ es el nivel de capacidad del examinado en una escala de calificación estándar, y $P(\theta)$ es la probabilidad de que una persona con nivel de capacidad θ conteste el reactivo correctamente. Suponiendo que $c = 0$, la fórmula 4.5 se reduce a la ecuación para el *modelo de dos parámetros*:

$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}}. \quad \textbf{(4.6)}$$

Otra suposición de que todos los reactivos son igualmente discriminantes produce la ecuación para el *modelo de un parámetro* o *modelo de Rasch*:

$$P(\theta) = \frac{1}{1 + e^{-1(\theta - b)}}. \quad \textbf{(4.7)}$$

Aunque el modelo de Rasch ha originado una gran cantidad de investigaciones psicométricas, el modelo de dos parámetros tiene por lo menos la misma popularidad.

Como se ilustra en la figura 4.2, la forma de una curva de respuesta a reactivos varía con los valores de los parámetros a y b. Ambas curvas de esta figura se construyeron con la función de dos parámetros de la fórmula 4.6. En la curva P, el parámetro de dificultad (b) es 1.00 y el parámetro de discriminación (a) es .5; en la curva Q, $b = .25$ y $a = .75$. Obsérvese que b es el valor de θ (el punto sobre el eje horizontal) que corresponde a $P(\theta) = .5$, y a es la pendiente de la curva en $P(\theta) = .5$. En el modelo de tres parámetros, b es el valor de $P(\theta)$ correspondiente a $.5(c + 1)$, donde c es el punto en que la curva de respuesta al reactivo cruza el eje vertical. Un ejercicio instructivo consiste en trazar varias curvas de respuesta a criterios de uno, dos y tres parámetros usando diversos valores de los parámetros adecuados. Las calificaciones del continuo de capacidad latente se expresan en unidades de calificación estándar (z), pero en la mayoría de las aplicaciones pedagógicas, las calificaciones z se transforman a una escala con una media de 300 y desviación estándar de 50.

En la práctica real, ni los parámetros de reactivos ni las calificaciones de capacidad latente (θ) de los examinados se conocen, y el problema es determinar la curva de respuesta a reactivos que mejor se ajuste a las respuestas a reactivos individuales. Esto incluye un procedimiento iterativo, de máxima aceptación, consistente en suponer ciertos valores iniciales para los parámetros de reactivos, calculando las $P(\theta)$ correspondientes a los diversos valores de θ, comparando las respuestas a reactivos pronosticadas con las reales y continuando con el proceso hasta alcanzar una solución más adecuada. El proceso de calcular parámetros de reactivo requiere de las respuestas de una gran cantidad de sujetos que son representativos de la población de exami-

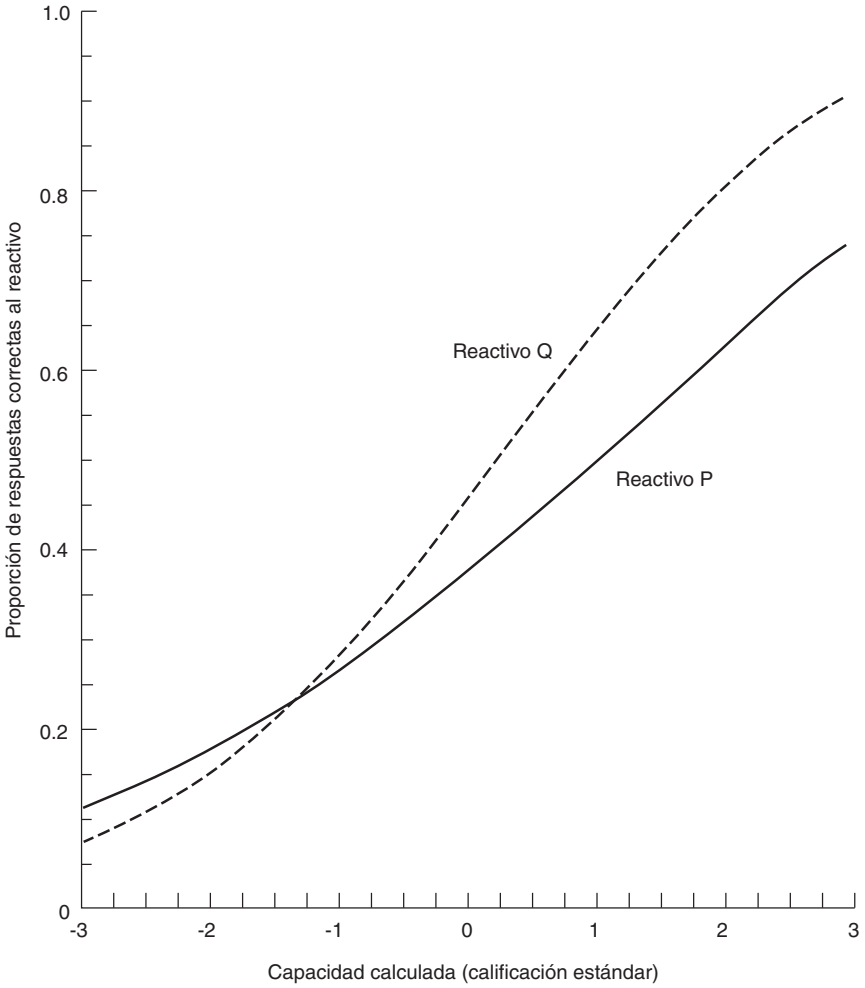


FIGURA 4.2 Dos curvas de respuesta a reactivos.
Vea la explicación en el texto.

nandos potenciales, aproximadamente 2,000 para el modelo de tres parámetros y 500 para el de un parámetro (Rasch).

Un rasgo importante de los parámetros de reactivos calculados es que son relativamente independientes del nivel de capacidad de la muestra particular de personas en que se basan. A diferencia de la metodología tradicional de evaluación, que confunde la dificultad y la discriminación de pruebas con la muestra de los individuos sometidos a la prueba, en la IRT estos parámetros son, al menos en teoría, independientes de la muestra particular de las personas evaluadas.

Además de proporcionar cálculos de parámetros de reactivos, la IRT puede usarse para estimar las calificaciones de los examinados en el continuo de capacidad latente. De hecho, este es el principal propósito de aplicar una prueba construida mediante los principios de la IRT.

Igual que al estimar parámetros de reactivos a partir de niveles de capacidad, el cálculo de calificaciones individuales en el continuo de capacidad latente es un proceso iterativo que se inicia al sustituir ciertos valores experimentales por la capacidad del examinado y los parámetros de reactivos supuestos en la ecuación logística apropiada. Las $P(\theta)$ resultantes se comparan entonces con las $P(\theta)$ reales, y el proceso continúa hasta que se obtiene una ecuación más adecuada. Los errores estándar de los valores estimados de θ, una medida de la variabilidad de las θ estimadas alrededor de las θ reales pero desconocidas, también pueden calcularse.

Otra propiedad interesante de la IRT, la invarianza de la capacidad del examinando con respuesta a los reactivos empleados para calcularla, se deriva del proceso de calcular las θ. Esta característica de la IRT significa que puede aplicarse una prueba de cualquier nivel de dificultad para determinar la posición de una persona en el continuo de capacidad latente. Sin embargo, el cálculo más preciso se obtiene cuando los reactivos que constituyen la prueba, y por ende la prueba misma, son los más adecuados, es decir, que están en el mismo nivel de dificultad que la capacidad del examinando.

La IRT se ha empleado para diversos propósitos, incluyendo la elaboración de pruebas, la calibración de calificaciones de pruebas con el fin de proporcionar un marco de referencia para interpretarlas, la estandarización de pruebas, la determinación del funcionamiento diferencial de reactivos (DIF), y evaluaciones adaptativas. Con respecto a la construcción de pruebas, pueden elegirse las áreas de la IRT sobre el continuo de capacidad donde se requieren mediciones más precisas, para que no se desperdicien reactivos en áreas menos importantes. Así, usando la IRT es posible desarrollar pruebas de clasificación, de selección y con referencias a criterios sumamente precisas, así como pruebas más tradicionales con un espectro amplio a lo largo del continuo de capacidad. El enfoque de la IRT sobre el DIF es trazar las curvas de respuesta a reactivos en forma separada para los dos o más grupos demográficos de personas (blancos contra negros, hombres contra mujeres, etc.). Las curvas de respuesta a reactivos con formas significativamente distintas para los grupos de comparación proporcionan pruebas del funcionamiento diferencial de los reactivos.

Una desventaja de la mayoría de los modelos de la IRT es la suposición de que un único rasgo latente subyace en la ejecución de las pruebas, pero los modelos multidimensionales han progresado. La mayoría de los modelos de la IRT se limita también a una calificación de 0-10, aunque también se han diseñado procedimientos más complejos que incluyen calificaciones de múltiples puntos, como en las escalas de calificación.

ESTANDARIZACIÓN Y NORMAS DE LAS PRUEBAS

Los datos sobre el desempeño de un grupo numeroso de individuos, como aquellos en quienes se basa el diseño de un instrumento, son útiles para propósitos de interpretación de calificaciones. Con el fin de cumplir esta tarea, deben estandarizarse la prueba, el inventario, la escala de clasificación y cualquier otro instrumento psicométrico.

Toda prueba estandarizada tiene instrucciones estándar de aplicación y calificación que deben seguirse estrictamente, sin dejar lugar a la interpretación personal o al sesgo. La estandarización también incluye aplicar la prueba a una muestra grande de personas (la *muestra de estandarización*) seleccionada como representante de la *población meta* a la que está destinada la prueba.

El principal propósito de estandarizar una prueba es determinar la distribución de puntuaciones crudas en la muestra de estandarización (*grupo norma*). Las calificaciones crudas obtenidas se transforman entonces en alguna forma de calificaciones derivadas o *normas*. Los principales tipos de normas son equivalentes de edad, de grado, rangos de percentilares y calificaciones estándar. La mayoría de los manuales de pruebas contiene tablas de normas con puntuaciones crudas y cierto tipo de calificaciones convertidas correspondientes. Así, la posición de

una persona en una prueba puede evaluarse con referencia a la tabla adecuada de normas y buscando los equivalentes de calificaciones convertidas de sus propias puntuaciones crudas. En este método de interpretación con referencias a normas, las normas obtenidas no funcionan como estándares del desempeño deseado, sino simplemente como un marco de referencia para interpretar calificaciones. Las normas indican la posición de una persona en la prueba con respecto a la distribución de las calificaciones obtenidas por personas de la misma edad cronológica, grado, sexo u otras características demográficas.

Al evaluar niños discapacitados, en ocasiones es preciso aplicar una *prueba fuera de nivel* diseñada para una edad o nivel de grado inferior al de la persona evaluada. Se requieren entonces normas especiales fuera de nivel para interpretar las calificaciones. Hay varias pruebas estandarizadas, como la Batería de Kaufman de Evaluación para Niños, que proporcionan evaluaciones fuera de nivel y las normas correspondientes.

En términos de tamaño de muestra y representatividad, con frecuencia las pruebas colectivas, y las de aprovechamiento en particular, se estandarizan de manera más adecuada que las pruebas individuales. Las normas para pruebas colectivas pueden estar basadas hasta en cien mil personas, mientras que el tamaño del grupo de norma para una prueba individual cuidadosamente estandarizada es más probable que sea de entre dos mil y cuatro mil. Sin embargo, una muestra de estandarización grande no garantiza que sea representativa de la población de interés. La muestra debe seleccionarse con sumo cuidado a fin de que sea representativa de la *población meta.*

Selección de una muestra de estandarización

Para funcionar con eficacia en la interpretación de calificaciones de pruebas, las normas deben ser apropiadas para el grupo o los individuos por evaluar. Por ejemplo, una calificación particular de un alumno de cuarto grado puede sobrepasar la del 80% de los niños de cuarto grado y la del 60% de los de sexto. Aunque puede ser de interés comparar la calificación de un estudiante de cuarto con las calificaciones de niños de tercero y sexto, la posición del alumno en su propio grupo (cuarto) es prioritaria. Siempre que se transforma una calificación con referencia a una tabla de normas, es importante tomar nota de las características de la muestra (edad, sexo, grupo étnico, educación, nivel socioeconómico, región geográfica) del grupo de norma en particular, e incluir esta información en todos los comunicados sobre el desempeño de la persona en las evaluaciones. Otra consideración importante es cuándo (en qué fecha) se obtuvieron las normas. En ciertas pruebas las normas pueden perder su vigencia en épocas de cambios sociales y educativos rápidos. Las modificaciones en el currículo escolar, por ejemplo, pueden requerir de una nueva estandarización o tal vez de modificar o reconstruir una prueba de aprovechamiento cada determinado número de años.

La forma en que una muestra de estandarización se selecciona de una población varía desde un muestreo aleatorio sencillo hasta estrategias más complejas, tales como el muestreo aleatorio estratificado y el muestreo por grupos. En un *muestreo aleatorio sencillo*, cada uno de los miembros de la población meta tiene la misma oportunidad de ser seleccionado. Empero, la aleatoriedad no garantiza que haya representatividad. En consecuencia, una forma más adecuada de estandarizar una prueba es empezar por categorizar, o *estratificar*, la población de una serie de variables demográficas (sexo, edad, nivel socioeconómico, región geográfica y similares) que presumiblemente están relacionadas con las calificaciones de la prueba. Entonces la cantidad de individuos seleccionados al azar de cada categoría o estrato es proporcional al número total de personas de la población que caen en ese estrato. Cuando se emplea este procedimiento de *muestreo aleatorio estratificado*, se reduce la probabilidad de elegir una muestra atípica o sesgada.

Las normas obtenidas de este modo proporcionan una base mejor para interpretar calificaciones de la prueba que las normas determinadas en una muestra aleatoria sencilla.

El *muestreo por grupos* es más económico que el muestreo aleatorio estratificado, y tiene mayores probabilidades de originar una muestra representativa de la población meta. El proceso se inicia al dividir una región geográfica designada o alguna otra entidad relevante en bloques o grupos. Entonces se elige al azar un porcentaje especificado de los grupos y dentro de cada uno se selecciona aleatoriamente una cantidad determinada de subunidades (escuelas, residencias, etc.). El último paso es administrar la prueba a cada persona de la subunidad, o por lo menos a una muestra aleatoria de personas con las características establecidas.

Administrar todos los reactivos de una prueba a una muestra aleatoria estratificada o a una muestra por grupo resulta tedioso y prolongado, por lo que se han propuesto estrategias menos costosas para obtener normas. Una de tales estrategias es elegir una muestra tanto de individuos como de reactivos. En el *muestreo de reactivos* se aplican distintas muestras de reactivos a muestras diferentes de personas seleccionadas al azar. Un grupo responde una serie de reactivos y otros grupos contestan otras series. El proceso es eficiente, en cuanto a que pueden aplicarse más reactivos a una gran cantidad de personas en un lapso bastante breve. Pueden realizarse entonces análisis de reactivos y determinarse normas basadas en calificaciones de muestras representativas para un amplio rango de contenidos de pruebas. Las normas derivadas del muestreo de reactivos son muy similares a las logradas mediante el procedimiento tradicional, pero más laborioso, de aplicar toda la prueba a una muestra representativa grande.

Las normas publicadas en manuales de pruebas son útiles para comparar la calificación de un examinado con las calificaciones de una muestra de personas de varias localidades, a veces una selección de todo el país. Pero en general los maestros están más interesados en saber cómo se desempeñaron los alumnos en comparación con otros en una escuela, un sistema escolar, estado o región particular, más que con las de una muestra de toda la nación. Cuando el interés se restringe a las calificaciones particulares de una escuela específica, el examinador querrá transformar las calificaciones crudas en *normas locales* mediante los procedimientos discutidos en las secciones subsiguientes. A menudo las normas locales se usan para fines de selección y colocación en escuelas y universidades.

Normas de edad y grado

Entre los tipos de normas más populares, sobre todo debido a que son bastantes fáciles de comprender para los usuarios, figuran las normas de edad y grado. Una *norma de edad* (equivalente de edad, edad educativa) es la calificación media de una prueba obtenida por las personas en una edad cronológica determinada; una *norma de grado* (equivalente de grado) es la calificación media obtenida por los estudiantes en un nivel de grado específico. Las normas de edad se expresan en doce intervalos de un mes que van, por ejemplo, para el décimo año, de 10 años, 0 meses, a 10 años, 11 meses. Las normas de grado se expresan en diez intervalos de un mes, con base en la suposición de que el crecimiento en la característica de interés durantes los meses de verano no tiene importancia. Por ejemplo, el rango de las normas de grado para el quinto grado es de 5-0 a 5-9, en intervalos de un mes desde el principio hasta el final del año escolar.

A pesar de su popularidad, las normas de edad y de grado tienen desventajas serias. El principal problema es que el progreso en las características cognoscitivas, psicomotoras o afectivas no es uniforme en todo el rango de edades o grados. Debido a que las unidades de edad y de grado se reducen progresivamente al aumentar la edad o el nivel de grado, una diferencia de evolución de dos meses en el aprovechamiento en el cuarto grado (por ejemplo, de 4-2 a 4-4) no es pedagógica-

mente equivalente a dos meses de evolución del aprovechamiento en un nivel de grado posterior (digamos, de 8-2 a 8-4). Las normas de edad y de grado implican erróneamente que la tasa de aumento de las capacidades evaluadas es constante de un año al siguiente, de modo que los especialistas en mediciones pedagógicas con frecuencia desaconsejan su uso. Se prefieren las normas en que la unidad de medida es menos variable a lo largo del rango de calificaciones.

Debido a su conveniencia, las normas de edad y de grado siguen usándose en el nivel escolar elemental o de primaria, donde las unidades de crecimiento son más constantes a lo largo del tiempo. No obstante, incluso en este nivel las normas de edad y de grado deben complementarse con normas de rangos percentilares o de calificaciones estándar para una edad o grado en particular.

Por lo común, los alumnos de un grado específico en el que se determinan normas de grado tienen un rango de edades bastante amplio: en las normas se incluyen las calificaciones de ciertos estudiantes que, de hecho, son mayores (o menores) que el alumno promedio en ese grado. Para proporcionar un índice más preciso de la calificación promedio de los alumnos en un nivel de grado establecido, en ocasiones se omiten las calificaciones de los estudiantes que son considerablemente mayores o menores que la edad modal, y la calificación media se calcula sólo en los estudiantes que tienen la edad apropiada para ese grado. Estas normas *restringidas* se conocen como *normas de edad modales*. Este tipo de normas, que casi no se encuentran en los manuales de pruebas de aprovechamiento contemporáneos, se mencionan aquí principalmente por su interés histórico.

Como se recordará, el término *edad mental* se mencionó en el breve análisis del capítulo 1 sobre la historia de la evaluación mental. Este concepto, que ideó Alfred Binet, es un tipo de norma de edad empleado en diversas pruebas de inteligencia. La calificación de edad mental de un examinado en particular corresponde a la edad cronológica del subgrupo de niños (todos de la misma edad cronológica) del grupo de estandarización cuya calificación media es la misma que la del examinando. Con fines pedagógicos, la práctica en muchas escuelas para evaluar a los retrasados mentales ha sido agruparlos de acuerdo con su edad mental en lugar de su edad cronológica.

Otra práctica de evaluación más antigua, que casi ha desaparecido, consiste en convertir las normas de edad en cocientes dividiendo las calificaciones de edad de cada examinando entre su edad cronológica (en meses) y multiplicando el cociente resultante por 100. El *cociente intelectual* (*relación de CI*) en la más antigua Escala de Inteligencia de Stanford-Binet, por ejemplo, se definió como:

$$\text{CI} = 100\left(\frac{\text{MA}}{\text{CA}}\right), \tag{4.8}$$

donde MA y CA son la edad mental y la edad cronológica del examinado en meses. De manera similar, un *cociente educativo* sobre ciertas pruebas de aprovechamiento se calculó como la relación entre la edad educativa (la norma de edad en una prueba de aprovechamiento educativo) y la edad cronológica en meses. Al comparar los resultados de un test de inteligencia con los de una prueba de aprovechamiento educativo, puede calcularse un *cociente de aprovechamiento* como la relación de la edad educativa con la edad mental. Algunos de estos cocientes aún se calculan evaluando las puntuaciones de pruebas, pero los especialistas en mediciones psicológicas no recomiendan esta práctica.

Normas percentilares

Las normas percentilares consisten en una tabla de percentiles que corresponden a puntuaciones crudas particulares. Las puntuaciones crudas se transforman como percentiles, y el porcentaje del grupo de norma inferior a una calificación en particular es el rango *percentilar* de dicha ca-

TABLA 4.2 Rangos percentilares y calificaciones estándar correspondientes a los puntos medios de una distribución de frecuencia de puntuaciones de pruebas

(1) INTERVALO DE CALIF.	(2) PUNTO MEDIO	(3) FRECUENCIA	(4) FRECUENCIA ACUMULATIVA	(5) RANGO PERCENTILAR DEL PUNTO MEDIO	(6) z	(7) Z	(8) z_n	(9) T	(10) NCE
750–799	774.5	3	248.5	99.4 (99)	2.59	76	2.51	75	103
700–749	724.5	11	241.5	96.6 (97)	2.03	70	1.82	68	88
650–699	674.5	18	227.0	90.8 (91)	1.48	65	1.33	63	78
600–649	624.5	27	204.5	81.8 (82)	.92	59	.91	59	69
550–599	574.5	49	166.5	66.6 (67)	.37	54	.43	54	59
500–549	524.5	65	109.5	43.8 (44)	–.19	48	–.16	48	47
450–499	474.5	38	58.0	23.2 (23)	–.74	43	–.73	43	35
400–449	424.5	25	26.5	10.6 (11)	–1.30	37	–1.25	38	24
350–399	374.5	13	7.5	3.0 (3)	–1.85	31	–1.88	31	11
300–349	324.5	1	.5	.2 (0)	–2.41	26	–2.88	21	–10

lificación. Las columnas 2 y 5 de la distribución que aparece en la tabla 4.2 muestran que, para este grupo de calificaciones, el rango percentilar de una calificación de 625 es aproximadamente 82, y el rango percentilar de una calificación de 475 es aproximadamente 23. Alternativamente, puede decirse que el octagésimo segundo percentil es 625 y el vigésimo tercero es 475.

Las normas percentilares a menudo se usan para fines de selección y colocación en una escuela o grado en particular, de manera que el procedimiento para calcular percentiles se describirá con cierto detalle. Las columnas 1 y 3 de la tabla 4.2 son una frecuencia de distribución de 250 calificaciones obtenidas en una prueba de capacidad académica, y la columna 2 da los puntos medios de los intervalos de calificaciones. Al calcular el valor inicial de la columna 4 (frecuencia acumulativa inferior al punto medio) para un intervalo en particular, se suman las frecuencias de todos los intervalos hasta ese intervalo. A este total se añade la mitad de la frecuencia de ese intervalo. Por ejemplo, el valor 227.0 para el intervalo 650-699 se calcula como $1 + 13 + 25 + 38 + 65 + 49 + 27 + \frac{1}{2}(18) = 227.0$. Dado que el valor inicial para un intervalo en particular de la columna 4 es la frecuencia acumulativa inferior al punto medio de ese intervalo, el rango percentilar de un punto medio de intervalo dado puede calcularse dividiendo la frecuencia acumulativa correspondiente de la columna 4 entre la cantidad total de calificaciones (n) y multiplicando el cociente resultante por 100. Para los datos de la tabla 4.2, $n = 250$, de modo que cada rango percentilar de la columna 5 es igual a 100 veces la frecuencia acumulativa correspondiente de la columna 4 dividida entre 250. Por ejemplo, el rango percentilar del punto medio 674.5 es $100(227/250) = 90.8 \approx 91$.

Los rangos percentilares son bastante fáciles de calcular y comprender, por lo que son más populares que las normas estándar de calificación. Las tablas de normas de rangos percentilares dentro de grupos de grados, edades cronológicas, género, ocupaciones, y otros grupos demográficos se incluyen en los manuales adjuntos a muchos instrumentos psicométricos. Desafortunadamente, el problema de las unidades de calificación desiguales, al que nos referimos antes en el análisis de las normas de edad y grado, no se resuelve con las normas de rangos percentilares.

Los rangos percentilares son medidas del nivel ordinal y no de intervalo (vea el apéndice A), y por lo tanto las unidades no son iguales en todas las partes de la escala. En relación con el atributo que se mide, la diferencia entre dos rangos percentilares ya sea en el extremo inferior o en el superior de la escala de Percentiles equivalentes (vea la figura 4.3) es mayor que la existente entre dos rangos percentilares con una diferencia numérica igual pero más cercana al centro de la escala.

El hecho de que las unidades de rangos percentilares se acumulen en la mitad y se dispersen en los extremos de la escala dificulta la interpretación de los cambios y las diferencias en estas calificaciones transformadas. Así, la diferencia de capacidad entre una persona con un rango percentilar de 5 y otra con uno de 10 en una prueba de aprovechamiento no es igual a la diferencia de capacidad entre una persona con un rango percentilar de 40 y otra que tenga uno de 45. En términos del atributo (habilidad) que se mide, la diferencia entre los rangos percentilares de 5 y 10, por ejemplo, es mayor que la existente entre los de 45 y 50; esto se debe a que es mayor la unidad de medida para la primera diferencia. Para interpretar normas de rangos percentilares en

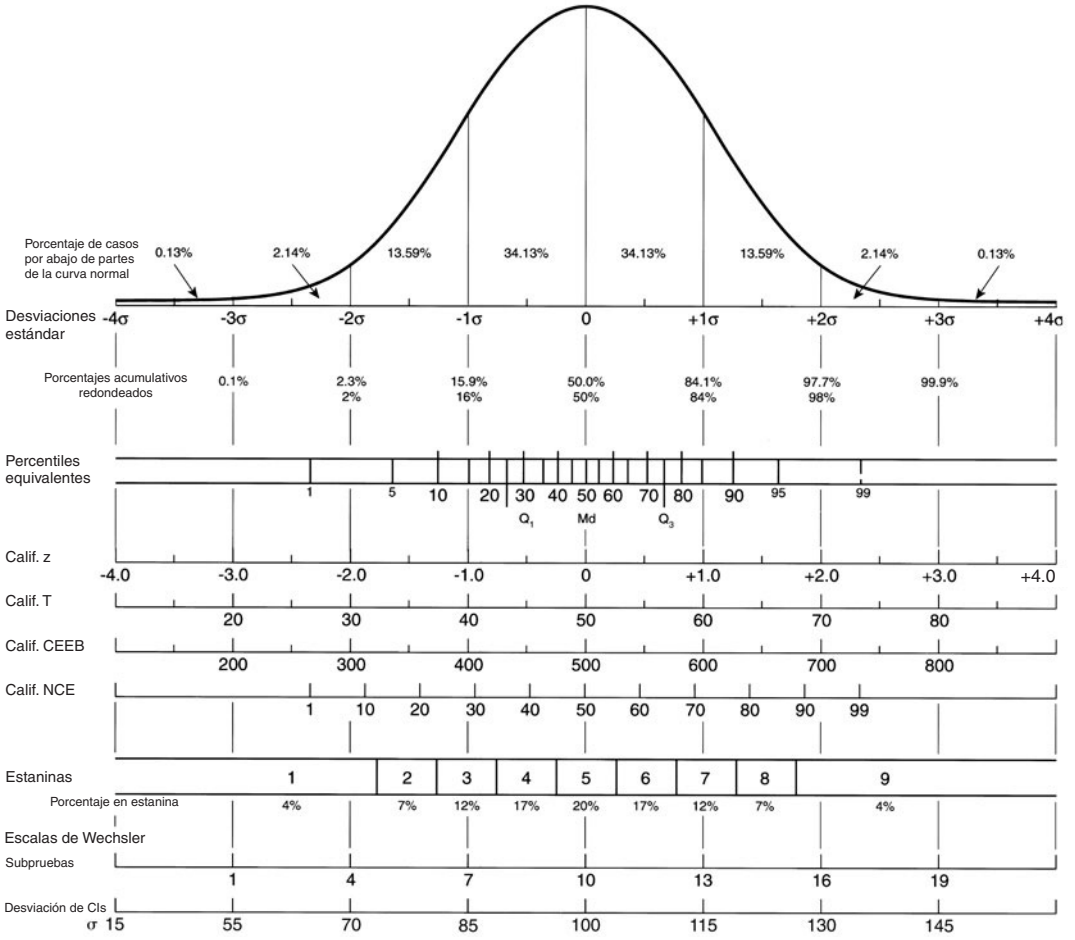


FIGURA 4.3 Rangos percentilares y calificaciones estándar correspondientes a varios puntos de la línea base de una distribución normal de calificaciones.

(H. G. Seashore, *Methods of expressing test scores*, The Psychological Corporation Test Service Bulletin, núm. 48, 1955.)

forma precisa, debemos recordar asignar un peso mayor a las diferencias de rango percentilar en los extremos que a las mismas diferencias cerca de la mitad de la escala.

Normas de calificación estándar

A diferencia de los rangos percentilares, las calificaciones estándar representan la medición en una escala de intervalos. Las *normas de calificación estándar* son puntuaciones convertidas que tienen cualesquier media y desviación estándar deseadas. Hay muchos tipos de calificaciones estándar, los cuales incluyen a las calificaciones *z*, *Z*, CEEB, de CI de desviación, estaninas, *T* y *NCE*.

Calificaciones *z* . Los equivalentes de calificaciones *z* de una distribución particular de puntuaciones crudas pueden determinarse como:

$$z = \frac{X - \overline{X}}{s}, \tag{4.9}$$

donde X es una puntuación cruda dada, $\overline{X}$ es la media aritmética, y s es la desviación estándar de las puntuaciones crudas. Transformar puntuaciones crudas en calificaciones *z* produce una distribución de calificaciones con la misma forma, pero con una media y desviación estándar distintas a la distribución de la puntuación cruda (X). La media de las calificaciones *z* es 0, y la desviación estándar es 1.

Las calificaciones *z* correspondientes a los puntos medios del intervalo incluidos en la columna 2 aparecen en la columna 6 de la tabla 4.2. La media y la desviación estándar de la distribución de calificaciones en la tabla 4.2 son 541.5 y 90.3, respectivamente. Por lo tanto, la calificación *z* correspondiente al punto medio 774.5 es (774.5 − 541.5)/90.3 = 2.58. Las calificaciones *z* correspondientes a los puntos medios de los demás intervalos pueden encontrarse del mismo modo. Las calificaciones *z* de varios puntos en la línea base de la curva normal se presentan en la figura 4.3.

Calificaciones Z. El hecho de que las calificaciones *z* pueden ser números decimales positivos o negativos origina cierta dificultad para manipularlas. El problema puede resolverse multiplicando las calificaciones *z* por una constante y añadiendo otra constante a los productos. Multiplicar *z* por 10, sumar 50 al producto, y redondear el resultado al número entero más cercano produce una calificación *Z*. La media de un conjunto de calificaciones *Z* es 50 y su desviación estándar es 10, pero la distribución de frecuencia de las calificaciones *Z* tiene la misma forma que la distribución original de las puntuaciones calificaciones crudas (vea la columna 7 de la tabla 4.2).

Calificaciones CEEB. En cierta época, las calificaciones estándar (calificaciones CEEB) sobre pruebas publicadas por el College Entrance Examination Board (Consejo de Evaluación de Ingreso a la Universidad) se determinaban multiplicando las calificaciones *z* correspondientes por 100 y sumando 500 a los productos. Por ejemplo, esto se hizo a las puntuaciones crudas de la Prueba de Aptitud Académica (SAT) aplicada en 1941, lo que produjo una nueva distribución con una media de 500 y desviación estándar de 100. Sin embargo, posteriormente las calificaciones obtenidas por estudiantes que se sometieron a la SAT no se transformaron de esta manera. Más bien, para garantizar una unidad de calificación constante por comparar los resultados de pruebas año con año, a partir de 1941 las calificaciones de la escala SAT se basaron en los resultados de la prueba aplicada ese año.[2]

[2]Las calificaciones estándar en la última versión de la SAT, renombrada como Prueba de Aptitud Académica, se basan en el desempeño de un millón de estudiantes que presentaron la prueba en 1994. Las nuevas calificaciones SAT se "reubicaron" para tener una media de 500 y desviación estándar de 100.

Calificaciones Wechsler. Las puntuaciones crudas en las subpruebas de las escalas de inteligencia de Wechsler se transformaron para tener una media de 10 y desviación estándar de 3. No obstante, las puntuaciones verbales, de ejecución y de escala total (CIs de desviación) en las pruebas de Wechsler se convirtieron a una distribución con una media de 100 y desviación estándar de 15 (vea las últimas dos líneas de la figura 4.3).

Calificaciones estándar normalizadas. Las normas de calificaciones estándar descritas arriba son simples transformaciones lineales de puntuaciones crudas. La media y las desviaciones estándar de las calificaciones transformadas son distintas de las de la distribución de la puntuación cruda, pero la forma de las dos distribuciones es idéntica. Si la distribución de la calificación es simétrica, también lo será la distribución de las calificaciones transformadas.

Para hacer las calificaciones de distintas pruebas más directamente comparables, se usa un procedimiento de transformación que no sólo afecte la media y la desviación estándar, sino que también cambie la forma de la distribución de las puntuaciones crudas a la de una distribución normal. Transformar un grupo de puntuaciones crudas en *calificaciones estándar normalizadas* empieza por calcular los rangos percentilares que corresponden a las puntuaciones crudas. Entonces, a partir de una tabla de áreas bajo la curva normal (apéndice B), se encuentra la calificación z que corresponde a cada rango percentilar. Por ejemplo, supóngase que los puntos medios (la columna 2) de la distribución de la tabla 4.2 deben convertirse en calificaciones estándar normalizadas. Debido a que los rangos percentilares de estos puntos medios ya se han encontrado (columna 5), empezamos por convertir los rangos percentilares en proporciones (por ejemplo, 99.4 se convierte en .994). Entonces, a partir de la tabla del apéndice B, se determinan las calificaciones Z bajo las cuales se encuentran las proporciones dadas del área. Así, la calificación z (z_n) bajo la cual se encuentra .994 del área bajo la curva es 2.51. Las otras calificaciones z normalizadas de la columna 8 de la tabla 4.2 se determinaron de manera similar. Para eliminar los puntos decimales y los números negativos, estas calificaciones z_n se transformaron en calificaciones T mediante la fórmula $T = 10z_n + 50$ (columna 9) y en calificaciones *NCE* (equivalente de curva normal) mediante la fórmula $NCE = 21z_n + 50$. Las calificaciones T van aproximadamente de 20 a 80 y las *NCE* de aproximadamente 0 a 100.

Las calificaciones z_n pueden transformarse en calificaciones normalizadas con cualesquier media y desviación estándar deseadas. Otra escala de calificación es la calificación *estanina* (nueve estándar) ejemplificada por la tercera escala desde abajo en la figura 4.3. En esta escala estándar normalizada, que tiene una media de 5 y desviación estándar de aproximadamente 2, hay nueve rangos distintos, o estaninas.[3] Estos rangos se designan con los números 1 al 9, y, como se muestra en la figura, cierto porcentaje de una distribución normal de pruebas cae dentro del intervalo representado por una estanina dada. Sin embargo, la calificación estanina no es una verdadera escala de calificaciones estándar, porque la primera y la novena estanina están abiertas. Obsérvese en la figura 4.3 que la amplitud de las estaninas 2 a 8 es igual, indican unidades de calificación estándar iguales, pero las estaninas 1 y 9 abarcan una distancia mucho más amplia.

Una ventaja de las calificaciones estaninas es que representan rangos más que puntos específicos. Esto contribuye a equilibrar la tendencia a considerar las calificaciones de pruebas como medidas precisas, invariables, de las diferencias individuales. Otro procedimiento que tiene el mismo efecto es registrar no sólo el rango percentilar o la calificación estándar correspondiente a una puntuacion cruda dada, sino también un rango percentilar o un intervalo de calificación estándar dentro de los cuales pueda esperarse razonablemente que caiga la verdadera posición del

[3]También se propusieron una calificación *sten* consistente en 10 unidades (Canfield, 1951) y una calificación *C* de 11 unidades (Guilford y Fruchter, 1973), pero sólo la segunda se ha usado en alguna medida.

examinado en la prueba. Esta práctica reconoce el hecho de que las calificaciones de las evaluaciones psicológicas y educativas no son exactas, sino que están sujetas a errores de medición.

IGUALACIÓN DE PRUEBAS

En muchas situaciones que implican la aplicación y la investigación de pruebas psicológicas, se requiere más de una versión de prueba. Las *formas paralelas* de una prueba son equivalentes en el sentido de que pueden contener los mismos tipos de reactivos de igual dificultad y que están altamente correlacionadas. Por lo tanto, las calificaciones que se obtienen en una forma son muy similares a las obtenidas por los mismos examinados en una segunda forma en el mismo nivel de edad o de grado que la primera forma. Desafortunadamente, elaborar pruebas paralelas es un proceso bastante caro y laborioso. Empieza con la preparación de dos pruebas, con el mismo tipo y número de reactivos, que originan las mismas medias y desviaciones estándar cuando se estandarizan en el mismo grupo de personas. Las formas paralelas producidas se igualan convirtiendo las calificaciones de una forma a las mismas unidades que las de la otra forma. Esto puede lograrse, por ejemplo, mediante el *método equipercentil* de cambiar las puntuaciones en cada forma a rangos percentilares. Entonces se prepara una tabla de calificaciones equivalentes sobre las dos formas equiparando el rango percentilar de *p* sobre la primera forma a la calificación del rango percentilar *p* sobre la segunda forma.

Al proceso de igualar, o más bien de hacer comparables, dos pruebas del mismo nivel de dificultad (por ejemplo, el mismo grado) se le conoce como *igualación horizontal*. Esto también puede realizarse *verticalmente*, como cuando se igualan las calificaciones de dos pruebas con distintos niveles de dificultad (grados diferentes). En general, el proceso de igualar incluye sujetar las pruebas a reactivos comunes o a un banco, como se realizó cada año con la Prueba de Aptitud Académica (SAT) estadounidense. Al usar un conjunto de reactivos en común que eran los mismos que un subconjunto de reactivos en por lo menos una forma anterior de la prueba, las calificaciones de cada forma nueva de la SAT que se aplicaba cada año se igualaban estadísticamente a formas previas en la prueba.

La *teoría de respuesta al ítem* (IRT), que prescribe métodos de calibración para un conjunto de reactivos de pruebas en un continuo de rasgos latente definidos de modo operativo (por lo común representados mediante calificaciones estándar en el eje horizontal de una curva de respuesta a ítemes), también se ha aplicado a la tarea de igualar pruebas. La propiedad de invarianza de la muestra en los parámetros de reactivos en la IRT, que se abordó en la explicación previa sobre análisis de reactivos, facilita el proceso de determinar calificaciones equivalentes o igualadas en distintas pruebas. El método de la IRT para igualar incluye buscar una ecuación lineal que transforme los parámetros del reactivo (índices de dificultad y de discriminación) de la versión de una prueba a los de una segunda versión. La metodología con que se establecen las constantes adecuadas para las ecuaciones lineales de transformación, de modo que los parámetros correspondientes en ambas pruebas se encuentren en la misma escala, se denomina *vinculación*. Los procedimientos de vinculación requieren que ambas pruebas compartan algunos reactivos en común (de soporte), o que un subconjunto de examinados resuelva ambas pruebas o un tercer examen que mida el mismo rasgo. Los procedimientos de igualación de la teoría de respuesta a los ítemes son económicos en cuanto a que también incluyen el muestreo de reactivos, en el que se aplican subconjuntos de reactivos seleccionados al azar a distintos grupos de personas seleccionadas también aleatoriamente.

Cualquiera que sea el método empleado para intentar igualar dos pruebas (equipercentil, de respuesta a ítemes, transformaciones de calificaciones lineales o no lineales), las pruebas que midan distintas características psicológicas o que tengan diferente confiabilidad no pueden, es-

trictamente hablando, igualarse. En casi todos los casos, lo mejor que puede hacerse es lograr que ambas pruebas o instrumentos psicométricos resulten "comparables".

RESUMEN

El principal objetivo de un análisis de reactivos es mejorar una prueba modificando o descartando los reactivos ineficaces. El análisis de reactivos también proporciona información específica sobre lo que saben o no los examinados. Los reactivos de pruebas pueden analizarse comparando respuestas a reactivos con calificaciones de criterio externo, como las notas asignadas por el maestro o las clasificaciones de los jefes, o de criterio interno, como calificaciones de prueba totales. Si el propósito es elaborar una prueba que pueda predecir al máximo las calificaciones con un criterio externo, entonces los reactivos deberían validarse contra el criterio.

Se calculan diversos análisis estadísticos para validar los reactivos de pruebas contra criterios externos e internos. Dichas estadísticas, que son índices de la relación entre reactivos calificados dicotómicamente (correcta-incorrecta) y calificaciones con la medida de criterio, constituyen una base para aceptar o rechazar reactivos específicos.

Dos sencillos coeficientes que pueden calcularse al analizar los reactivos de una prueba elaborada por maestros son el índice de dificultad de reactivos (p) y el índice de discriminación de reactivos (D). Estos índices se aplican a reactivos tanto con referencias a normas como con referencias a criterios. El valor óptimo de p depende de los propósitos de la prueba y de la cantidad de opciones por reactivo. En la mayoría de los casos se requiere un valor D de .30 o mayor para que un reactivo sea aceptable.

Además de calcular los índices de dificultad y de discriminación de los reactivos de prueba, los reactivos deben examinarse en cuanto a sesgos, ambigüedad y los efectos de la velocidad. Las variaciones marcadas de la uniformidad en la distribución de frecuencia de las respuestas a los distractores son un signo de deficiencias en el funcionamiento del reactivo.

Al elaborar una curva característica de los reactivos, la proporción de examinados que dan la respuesta en clave a un reactivo se traza contra las calificaciones con un criterio interno (calificaciones de prueba totales) o externo. Una extensión del método de curva característica de los reactivos, conocida como teoría de respuesta a los ítemes, conlleva incluir parámetros de dificultad, discriminación y adivinanza en una ecuación logística, o bien derivar valores de estos parámetros para dicha ecuación. La ecuación logística relaciona la proporción de examinados que contestaron el reactivo de manera correcta con cálculos de sus calificaciones en un continuo específico de capacidad u otra característica unidimensional.

La estandarización consiste en aplicar una prueba a una muestra representativa de personas en condiciones estándar (uniformes) y mediante un procedimiento estándar. Las normas calculadas a partir de las puntuaciones de prueba obtenidas conforman un marco de referencia para interpretar puntuaciones alcanzadas por personas que después se someten a la prueba. Tradicionalmente, las normas se han establecido evaluando una muestra (aleatoria, aleatoria estratificada, por grupo) de la población para la que está destinada la prueba. De menor costo y más eficientes que los procedimientos convencionales de estandarización de pruebas son las técnicas de muestreo de reactivos, en las que se toman muestras no sólo de las personas sino también de los reactivos y distintos grupos de examinados responden diferentes conjuntos de reactivos.

Dependiendo de las necesidades y recursos de los usuarios de pruebas, las normas pueden calcularse en muestras locales, regionales o nacionales. Las normas de edad y grado, que se establecen con mayor frecuencia para pruebas de aprovechamiento, permiten comparar calificaciones de pruebas individuales con el promedio de calificaciones de niños de cierta edad o grado. La principal desventaja de las normas de edad y grado es que el progreso en el aprovechamiento o capacidad no es uniforme a través de la edad o los niveles de grado. Las normas de

rangos percentilares, en las que las puntuaciones crudas de una prueba se convierten en porcentajes de personas en el grupo de estandarización que alcanzaron esas calificaciones o menos, también se ven afectadas por el problema de desigualdad en las unidades de calificación. Las normas de rango percentilar, de edad y de grado son bastante fáciles de comprender y su uso es conveniente; por ello, sin duda continuarán siendo populares.

Las normas de calificaciones estándar se convierten en calificaciones que tienen una media y una desviación estándar designadas. A diferencia de las medidas ordinales representadas por la edad, el grado y las normas de rango percentilar, las calificaciones estándar (z, T, CEEB y otras) son medidas de nivel de intervalo. No todas las calificaciones estándar se distribuyen normalmente, pero pueden convertirse con facilidad en calificaciones estándar normalizadas.

Las calificaciones de pruebas paralelas pueden escalarse para lograr igualdad, si no se igualan estrictamente, de varias maneras. Tradicionalmente, las pruebas se han igualado por el método equipercentil, pero los métodos más recientes acarrean modelos de respuesta a reactivos técnicamente más complejos.

PREGUNTAS Y ACTIVIDADES

1. ¿Cuáles son los índices de dificultad (p) y de discriminación (D) de una prueba administrada a 75 personas si 18 de las del grupo superior (27% superior en el total de calificaciones de la prueba) y 12 del grupo inferior (27% inferior del total de calificaciones de la prueba) aciertan en el reactivo? Obsérvese que el redondeo da como resultado 20 personas en el grupo superior y 20 en el grupo inferior.

2. Calcule los índices de dificultad (p) y de discriminación (D) de un reactivo de una prueba con referencia a criterio aplicada a 50 personas, 30 de las cuales obtuvieron calificaciones en el nivel del criterio o superior, y 20 consiguieron calificaciones por debajo del nivel de criterio. De quienes alcanzaron o superaron el nivel del criterio, 20 acertaron en el reactivo; entre las que quedaron bajo el nivel del criterio, 10 dieron la respuesta correcta al reactivo.

3. La siguiente tabla en dos direcciones indica si cada una de las 20 personas acertó (a) o falló (f) en cada uno de los 10 reactivos en una prueba de opción múltiple con cuatro opciones. Clasificando a los examinandos de la A a la J en el grupo superior y de la K a la T en el grupo inferior sobre la puntuación total de la prueba (vea la última línea de la tabla), calcule los índices de dificultad y de discriminación para cada reactivo. Escriba estos valores en las últimas dos columnas de la tabla. Al examinar los índices p y D, decida qué reactivos son aceptables y cuáles necesitan modificarse o descartarse.

Examinado

Reactivo	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	*p*	*D*
1	a	a	a	f	f	f	a	a	a	a	f	a	a	f	f	f	a	f	f	f		
2	a	a	f	a	f	a	f	a	f	a	a	f	f	a	f	f	f	f	a	f		
3	a	f	a	a	a	f	a	f	a	f	f	a	a	f	f	a	f	f	f	f		
4	a	a	a	a	a	a	f	a	f	a	a	f	f	a	a	f	f	f	f	f		
5	a	a	f	a	a	a	a	a	f	f	f	a	f	f	f	f	f	f	f	f		
6	a	a	a	a	a	a	f	a	a	a	a	f	a	a	f	a	f	a	f	a		
7	a	f	f	f	a	a	a	a	f	a	f	f	a	f	a	f	f	a	a	f		
8	a	a	a	a	a	f	f	f	a	f	a	a	f	f	f	a	a	f	f	f		
9	a	a	a	a	f	a	a	f	a	a	a	a	f	a	a	f	f	f	f	f		
10	a	a	a	f	a	a	a	f	a	f	f	f	f	f	f	f	a	f	f	f		
Calif.	10	8	7	7	7	7	6	6	6	6	5	5	4	4	3	3	3	2	2	1		

4. Suponga que Jorge obtiene una puntuación cruda de 65 en una prueba aritmética con una media de 50 y desviación estándar de 10, pero obtiene una puntuación cruda de 80 en una prueba de lectura con una media de 75 y desviación estándar de 15. ¿Cuáles son las calificaciones z y Z en las pruebas? ¿Jorge es mejor en aritmética o en lectura?

5. Con referencia a la tabla de áreas bajo la curva normal (apéndice B), busque las calificaciones z correspondientes a los rangos percentilares 10°, 20°, 30°, 40°, 50°, 60°, 70°, 80° y 90°. Luego convierta las calificaciones z en calificaciones T, CEEB, NCE y estaninas.

6. Construya una distribución de frecuencia a partir de las 30 calificaciones que aparecen enseguida, use un ancho de intervalo de 3. Luego calcule el rango percentilar y las calificaciones z, Z, z_n y T correspondientes a los puntos medios del intervalo.

82	85	70	91	75	88	78	82	95	79
86	90	87	77	87	73	80	96	86	81
85	93	83	89	92	89	84	83	79	74

7. ¿Por qué las normas de calificaciones estándar se consideran superiores a las normas de edad, de grado y de rango percentilares?

8. A continuación se presenta una lista de calificaciones de una prueba de semejanzas de ocho reactivos en la que las posibles calificaciones van de 0 a 16. Calcule el rango percentilar, la calificación z, y la calificación T correspondiente a cada una de las puntuaciones crudas. Consulte el apéndice A como ayuda.

CALIF. CRUDA	FRECUENCIA	RANGO PERCENTILAR	z	Z	T
16	8				
15	26				
14	71				
13	140				
12	171				
11	223				
10	272				
9	250				
8	257				
7	209				
6	183				
5	124				
4	89				
3	79				
2	51				
1	23				
0	25				

9. Describa los procedimientos para igualar (es decir, volver comparables) las calificaciones en dos pruebas diseñadas como formas paralelas.

CAPÍTULO CINCO

CONFIABILIDAD Y VALIDEZ

La estandarización es un paso importante en el diseño y la evaluación de pruebas psicológicas y otros instrumentos de evaluación, pero no es el último paso. Antes de que una prueba pueda utilizarse con cierta seguridad, debe obtenerse información acerca de su confiabilidad y validez por lo que a sus propósitos específicos concierne.

CONFIABILIDAD

Ningún instrumento psicométrico puede considerarse de valor a menos que sea una medida consistente, o *confiable*. En consecuencia, una de las primeras cosas que será necesario determinar acerca de una prueba de elaboración reciente es si resulta lo suficientemente confiable como para medir lo que fue diseñada para medir. Si, en ausencia de cualquier cambio permanente en una persona debido al crecimiento, al aprendizaje, a alguna enfermedad o lesión, las puntuaciones en una prueba varían con la ocasión o la situación, es probable que la prueba no sea lo suficientemente confiable como para ser usada en describir y evaluar a la gente y hacer predicciones sobre su conducta. Hablando en términos estrictos, más que ser una característica de una prueba, la confiabilidad es una propiedad de las puntuaciones obtenidas cuando se administra la prueba a un grupo particular de personas en una ocasión particular y bajo condiciones específicas (Thompson, 1994).

Note que *confiabilidad* no es lo mismo que *estabilidad:* al determinar la confiabilidad se asume que la prueba mide una característica relativamente estable. A diferencia de la inestabilidad, la falta de confiabilidad es resultado de errores de medición producidos por estados internos temporales, como la baja motivación o la falta de disposición, o de condiciones externas como un ambiente de prueba incómodo o con distracciones.

Teoría clásica de la confiabilidad

En la teoría clásica de los tests se supone que la calificación observada de una persona en una prueba está compuesta por una puntuación "real" más algún error no sistemático de medición. La *puntuación real* de una persona en una prueba particular se define como el promedio de las puntuaciones que obtendría si presentara la prueba un número infinito de veces. Es obvio que la puntuación real de una persona nunca puede medirse de manera exacta; tiene que ser estimada

a partir de su puntuación observada en la prueba. También se asume en la teoría clásica de los tests que la varianza de las puntuaciones observadas (s^2_{obs}) de un grupo de personas es igual a la varianza de sus puntuaciones reales (s^2_{rea}) más la varianza debida a los errores no sistemáticos de medición (s^2_{err}):

$$s^2_{obs} = s^2_{rea} + s^2_{err} \quad \textbf{(5.1)}$$

Entonces la confiabilidad (r_{11}) de las calificaciones se define como la razón de la varianza de la calificación real con la varianza de la calificación observada, o la proporción de la varianza observada que es explicada por la varianza real:

$$r_{11} = \frac{s^2_{tru}}{s^2_{obs}} \quad \textbf{(5.2)}$$

La proporción de la varianza observada explicada por la varianza de error o que no se explica por la varianza real puede determinarse a partir de las fórmulas 5.1 y 5.2 como:

$$\frac{s^2_{err}}{s^2_{obs}} = 1 - r_{11}. \quad \textbf{(5.3)}$$

La confiabilidad de un conjunto de calificaciones en una prueba se expresa como un número decimal positivo que fluctúa entre .00 y 1.00. Una r_{11} de 1.00 indica una confiabilidad perfecta, y una r_{11} de .00 indica una falta absoluta de confiabilidad de la medición. Como la varianza de las calificaciones reales no puede calcularse de manera directa, la confiabilidad se estima analizando los efectos de variaciones en las condiciones de la administración y el contenido de la prueba en las calificaciones observadas. Como advertimos antes, la confiabilidad no es influida por cambios sistemáticos en las calificaciones que tienen un efecto similar en todos los examinados, sino sólo por cambios no sistemáticos que tienen efectos diferentes en personas distintas. Dichos factores no sistemáticos influyen en la varianza de error y, por lo tanto, en la confiabilidad de las calificaciones en la prueba. Cada uno de los diversos métodos para estimar la confiabilidad (test-retest, formas paralelas, consistencia interna) toma en consideración los efectos de circunstancias algo diferentes que pueden producir cambios no sistemáticos en las puntuaciones y, por ende, afectan la varianza de error y el coeficiente de confiabilidad.

Coeficiente test-retest

Se calcula un *coeficiente test-retest* para determinar si un instrumento mide de manera consistente de una ocasión a otra. Este coeficiente, conocido también como *coeficiente de estabilidad*, se encuentra correlacionando las calificaciones obtenidas por un grupo de personas en una aplicación con sus puntuaciones en la segunda aplicación de la prueba. El procedimiento test-retest toma en consideración los errores de medición que resultan de diferencias en las condiciones (ambientales, personales) asociadas con las dos ocasiones en que se administró la prueba. Dado que en ambas ocasiones se aplicó la misma prueba, los errores debidos a diferentes muestras de los reactivos de la prueba no se reflejan en un coeficiente test-retest. Además, es probable que las diferencias entre las condiciones de la aplicación sean mayores luego de un intervalo largo

que de uno corto. Como resultado, la magnitud de un coeficiente de confiabilidad test-retest tiende a ser mayor cuando el intervalo entre la prueba inicial y el retest es corto (unos cuantos días o semanas) que cuando es largo (meses o años).

Coeficiente de formas paralelas

Cuando el intervalo entre la prueba inicial y el retest es corto, los examinados recuerdan, por lo general, muchas de las preguntas y respuestas de la prueba inicial. Como es obvio, esto afecta sus respuestas en la segunda aplicación, un hecho que por sí mismo no cambia el coeficiente de confiabilidad si todos recuerdan igual cantidad. Sin embargo, por lo regular algunas personas recuerdan más del material de la prueba que otras, ocasionando que la correlación entre el test y el retest sea menos que perfecta. Lo que parece necesitarse para superar esta fuente de error es una forma paralela del instrumento, esto es, una que conste de reactivos similares pero no de los mismos reactivos. Entonces puede calcularse como índice de confiabilidad un *coeficiente de formas paralelas,* también conocido como *coeficiente de equivalencia*.

En principio, la idea de formas paralelas es razonable: al aplicar una forma paralela luego de un intervalo apropiado que sigue a la aplicación de la primera forma puede determinarse un coeficiente de confiabilidad que refleje los errores de medición debidos a los diferentes reactivos y los distintos momentos de aplicación. Para controlar los efectos de confusión de la forma de la prueba con el momento de la aplicación, la forma A debe administrarse primero a la mitad del grupo y la forma B a la otra mitad; luego, en la segunda aplicación, el primer grupo presenta la forma B y el segundo la forma A. La correlación resultante entre las calificaciones de las dos formas, conocida como *coeficiente de estabilidad y equivalencia*, toma en cuenta errores debidos a los diferentes momentos de aplicación o a los distintos reactivos.

Coeficientes de consistencia interna

Se dispone de formas paralelas para una serie de pruebas, en particular para pruebas de habilidad (aprovechamiento, inteligencia, aptitudes especiales). Sin embargo, una forma paralela de una prueba a menudo es costosa y difícil de elaborar. Por esta razón se elaboró un método menos directo de tomar en cuenta los efectos de diferentes muestras de los reactivos de una prueba sobre la confiabilidad. Éste es el *método de consistencia interna*, que incluye el método de división por mitades de Spearman, las fórmulas de Kuder-Richardson y el coeficiente alfa de Cronbach. Sin embargo, los errores de medición causados por diferentes condiciones o momentos de aplicación no se reflejan en un coeficiente de consistencia interna. En consecuencia, este tipo de coeficientes no pueden verse como verdaderos equivalentes de los coeficientes test-retest o de formas paralelas.

Método de división por mitades. En este enfoque simplificado de la consistencia interna una sola prueba se considera compuesta por dos partes (formas paralelas) que miden la misma cosa. De este modo, puede aplicarse una prueba y asignar calificaciones separadas a sus dos mitades seleccionadas de manera arbitraria. Por ejemplo, los reactivos con números nones pueden calificarse por separado de los que tienen números pares. Entonces la correlación (r_{oe}) entre los dos conjuntos de calificaciones obtenidas por un grupo de personas es un coeficiente de confiabilidad de formas paralelas para una mitad de la prueba tan larga como la prueba original. Suponien-

do que las dos mitades equivalentes tienen medias y varianzas iguales, la confiabilidad de la prueba como un todo puede estimarse mediante la *fórmula Spearman-Brown*:

$$r_{11} = \frac{2r_{oe}}{1+r_{oe}} \tag{5.4}$$

Para demostrar el uso de la fórmula 5.4, suponga que la correlación entre las calificaciones totales obtenidas en los reactivos con números nones y en los reactivos con números pares de una prueba es .80. Entonces la confiabilidad estimada de toda la prueba es $r_{11} = 2(.80)/(1 + .80) = .89$.

Método de Kuder-Richardson. Una prueba puede dividirse de muchas formas diferentes en dos mitades que contengan igual número de reactivos. Como cada forma puede dar por resultado un valor algo diferente de r_{11}, no queda claro qué estrategia de división producirá el mejor estimado de confiabilidad. Una solución al problema es calcular el promedio de los coeficientes de confiabilidad obtenidos de todas las divisiones por mitades como el estimado global de confiabilidad. Esto puede hacerse, pero el siguiente procedimiento abreviado fue elaborado por Kuder y Richardson (1937).

Bajo ciertas condiciones, la media de todos los coeficientes de división por mitades puede estimarse mediante una de las siguientes fórmulas:

$$r_{11} = \frac{k[1 - \Sigma p_i(1 - p_i)/s^2]}{k - 1} \tag{5.5}$$

$$r_{11} = \frac{k - \overline{X}(k - \overline{X})/s^2}{k - 1} \tag{5.6}$$

En estas fórmulas, k es el número de reactivos en la prueba, $\overline{x}$ es la media de las calificaciones totales de la prueba, s^2 es la varianza de las calificaciones totales de la prueba (calculadas con n en lugar de $n - 1$ en el denominador), y p_i es la proporción de examinados que dan la respuesta de la clave al reactivo i. Las p_i se suman a lo largo de todos los reactivos k. Las fórmulas 5.5 y 5.6 se conocen como fórmulas Kuder-Richardson (K-R) 20 y 21, respectivamente. A diferencia de la fórmula 5.5, la 5.6 se basa en la suposición de que todos los reactivos son de igual dificultad; esto también conduce a una estimación más conservadora de la confiabilidad y es más fácil de calcular que la fórmula 5.5.

Para demostrar la aplicación de la fórmula 5.6, suponga que una prueba que contiene 75 reactivos tiene una media de 50 y una varianza de 100. Entonces $r_{11} = [75 - 50(75 - 50)/100]/74 = .84$.

Coeficiente alfa. Las fórmulas 5.5 y 5.6 son casos especiales del coeficiente alfa más general (Cronbach, 1951). El *coeficiente alfa* se define como

$$\alpha = \frac{k(1 - \Sigma s_i^2/s_t^2)}{k - 1} \tag{5.7}$$

donde k es el número de reactivos, s_i^2 la varianza de las calificaciones en el reactivo i, y s_t^2 la varianza de las calificaciones totales de la prueba. Las fórmulas de Kuder-Richardson sólo son aplicables cuando los reactivos de la prueba se califican con 0 o 1, pero el coeficiente alfa es una fórmula general para estimar la confiabilidad de una prueba que consta de reactivos en los cuales pueden asignarse calificaciones de distinto peso a respuestas diferentes.

Todos los procedimientos de consistencia interna (división por mitades, Kuder-Richardson, coeficiente alfa) sobrestiman la confiabilidad de las pruebas de velocidad. En consecuencia,

deben modificarse para proporcionar estimaciones razonables de confiabilidad cuando la mayoría de los examinados no termina la prueba en el tiempo permitido. Para ello, una posibilidad consiste en aplicar las dos mitades de la prueba en momentos diferentes, pero con límites de tiempo iguales. Se calcula entonces la correlación entre las calificaciones de las dos mitades cronometradas por separado y los coeficientes resultantes se corrigen con la fórmula 5.4. También pueden usarse los procedimientos de test-retest y de formas paralelas para estimar las confiabilidades de las pruebas de velocidad.

Confiabilidad entre calificadores

Salvo por errores administrativos, las calificaciones calculadas por dos calificadores diferentes de una prueba objetiva presentada por un individuo deben ser idénticas. Sin embargo, la calificación de las pruebas de ensayo y orales, además de otros juicios evaluativos (calificaciones de personalidad, calificación de pruebas proyectivas) es un proceso bastante subjetivo. Al evaluar las calificaciones que implican el juicio subjetivo del calificador, es importante conocer el grado en que diferentes calificadores están de acuerdo en las calificaciones y otros valores numéricos dados a las respuestas de diferentes examinados y reactivos. El enfoque más común para determinar la *confiabilidad entre calificadores* es hacer que dos personas califiquen las respuestas de un número considerable de examinados y calcular luego la correlación entre los dos conjuntos de calificaciones. Otro enfoque es hacer que muchas personas califiquen las respuestas de un examinado o, mejor aún, que muchas personas califiquen las respuestas de varios examinados. Esta última estrategia arroja un *coeficiente intraclase* o *coeficiente de concordancia*, el cual es un coeficiente generalizado de confiabilidad entre calificadores. En muchos libros de estadística se describen los procedimientos para calcular estos coeficientes.

Las pruebas orales no se distinguen por tener una elevada confiabilidad, pero se dispone de formas especiales que pueden mejorar la objetividad, y por ende la confiabilidad, con la que se juzga el desempeño oral (vea la forma 3.1 en la página 58). Aunque los exámenes orales tienen, por lo general, una confiabilidad menor que pruebas escritas comparables, la atención cuidadosa al diseño de las preguntas orales, a la elaboración de las respuestas modelo a las preguntas antes de aplicar la prueba, y al uso de calificadores múltiples, puede mejorar la confiabilidad de las calificaciones en las pruebas orales. Dichos procedimientos han dado por resultado coeficientes de confiabilidad entre calificadores de .60 y .70 para las pruebas orales aplicadas en ciertos cursos de licenciatura, posgrado y de escuelas profesionales. Otras sugerencias para mejorar la confiabilidad de las evaluaciones del desempeño oral incluyen alentar a los examinados a demorar la respuesta hasta que hayan pensado por un momento en la pregunta, y registrar las respuestas de manera electrónica para que más tarde sean reproducidas y reevaluadas por los calificadores.

Interpretación de los coeficientes de confiabilidad

Los coeficientes de confiabilidad de instrumentos afectivos como las listas de verificación, escalas de calificación e inventarios de personalidad, intereses o actitudes, por lo general son más bajos que los de las pruebas cognitivas de aprovechamiento, inteligencia o habilidades especiales. Sin embargo, los coeficientes de confiabilidad obtenidos con esos instrumentos afectivos pueden ser bastante respetables, y los obtenidos con los instrumentos cognitivos en ocasiones son bastante bajos.

¿Qué tan alto debe ser un coeficiente de confiabilidad para que una prueba u otro instrumento psicométrico sean útiles? La respuesta depende de lo que planeemos hacer con las puntuaciones de la prueba. Cuando una prueba va a utilizarse para determinar si las calificaciones promedio de dos grupos de personas son significativamente diferentes, un coeficiente de confia-

bilidad de .60 a .70 puede ser satisfactorio. Por otro lado, cuando se utiliza la prueba para comparar la calificación de una persona con la de otra, o la calificación de una persona en una prueba con su calificación en otro instrumento, se necesita un coeficiente de confiabilidad de al menos .85 para determinar si diferencias pequeñas en las calificaciones son significativas.

Variabilidad y extensión de la prueba

Como con otras medidas de relación, los coeficientes de confiabilidad tienden a ser más altos cuando la varianza de las puntuaciones de la prueba, las puntuaciones del reactivo, las calificaciones u otras variables que son evaluadas, es grande que cuando es pequeña. Como la varianza de la calificación de la prueba se relaciona con la extensión de ésta, un método para incrementar la confiabilidad es hacer la prueba más larga. Sin embargo, la simple inclusión de más reactivos en una prueba no necesariamente incrementa su confiabilidad. Los nuevos reactivos deben ser del mismo tipo general y medir la misma cosa que los reactivos que ya contiene la prueba. De hecho, agregar reactivos que miden algo diferente de lo que miden los reactivos originales puede dar lugar a una reducción en la confiabilidad.

La fórmula general de Spearman-Brown es una expresión del efecto que tiene sobre la confiabilidad el alargar una prueba incluyendo más reactivos del mismo tipo general. Esta fórmula, una generalización de la fórmula 5.4, es:

$$r_{mm} = \frac{mr_{11}}{1 + (m - 1)r_{11}} \tag{5.8}$$

donde m es el factor por el cual se alarga la prueba, r_{11} la confiabilidad de la prueba original no alargada, y r_{mm} la confiabilidad estimada de la prueba alargada. Por ejemplo, si una prueba de 20 reactivos que tiene un coeficiente de confiabilidad de .70 se hace tres veces más larga agregando 40 reactivos más, la confiabilidad estimada de la prueba alargada será $3(.70)/[1 + 2(.70)] = .875$. La figura 5.1 ilustra los efectos que produce sobre la confiabilidad el incrementar el número de reactivos en una prueba por un factor de $1\frac{1}{2}$, 2, 3, 4 o 5. Note que el incremento creciente en la confiabilidad es menor cuando la confiabilidad inicial es alta y con incrementos sucesivamente mayores en la extensión de la prueba.

Resolver la fórmula 5.8 para m arroja la siguiente fórmula para determinar cuántas veces más extensa debe ser una prueba de confiabilidad r_{11} a fin de obtener una confiabilidad deseada (r_{11}):

$$m = \frac{r_{mm}(1 - r_{11})}{r_{11}(1 - r_{mm})} \tag{5.9}$$

Esta fórmula puede utilizarse para determinar el incremento necesario en la longitud de la prueba y, en consecuencia, el número de reactivos que deben agregarse para incrementar la confiabilidad de un valor desde r_{ll} hasta r_{mm}.

Además de depender del número de reactivos, la varianza y la confiabilidad de una prueba son afectadas por la heterogeneidad de la muestra de personas que la presentan. Entre mayor sea el rango de diferencias individuales en cierta característica, mayor será la varianza de las calificaciones en una medida de esa característica. En consecuencia, el coeficiente de confiabilidad de una prueba u otro instrumento de evaluación será mayor en un grupo más heterogéneo con una varianza más grande en la calificación de la prueba. El que la confiabilidad de una prueba varíe con la naturaleza del grupo probado se refleja en la práctica de informar acerca de coeficientes de confiabilidad separados para grupos que difieren en edad, grado, género y posición

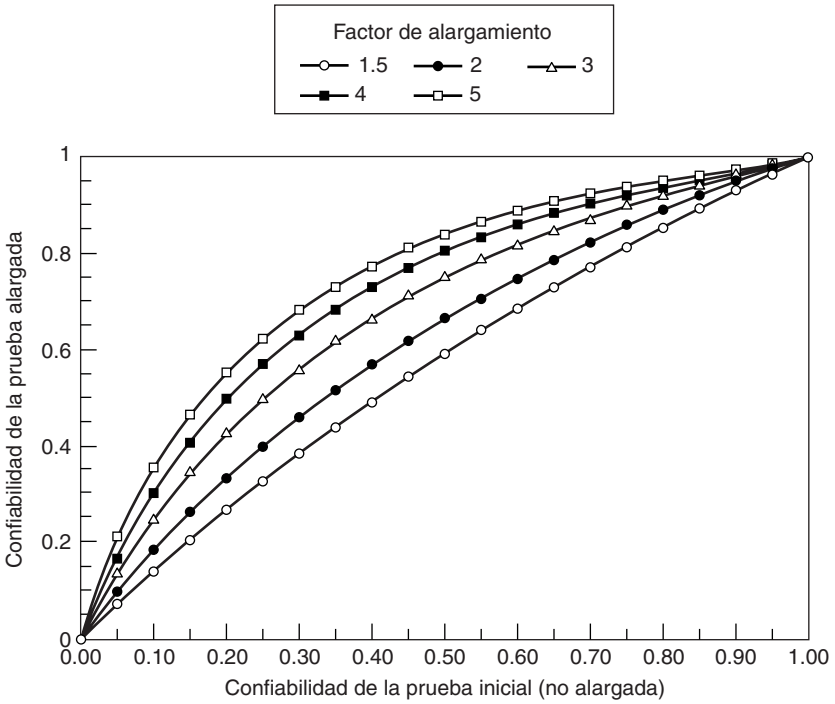


Figura 5.1 Confiabilidad de una prueba alargada como función de la confiabilidad inicial y el factor de alargamiento. La confiabilidad se incrementa a medida que se agregan a una prueba más reactivos del mismo tipo general, pero el monto del incremento es mayor cuando la confiabilidad inicial es baja. Además, la confiabilidad de la prueba alargada se nivela gradualmente conforme la prueba se vuelve cada vez más larga.

socioeconómica. La asociación entre la varianza y la confiabilidad de una prueba también se advierte en que las pruebas compuestas, sobre todo por reactivos de dificultad intermedia (valores *p* de alrededor de .50), tienden a ser más confiables que las pruebas donde la mayoría de los reactivos tienen índices más altos o más bajos de dificultad.

Error estándar de medición

Puesto que se desconoce la varianza de las calificaciones reales, no puede calcularse la confiabilidad de manera directa a partir de la fórmula 5.2. Sin embargo, dado un estimado de la confiabilidad, puede calcularse la varianza de la calificación real a partir de la fórmula 5.2 o, lo que es de mayor interés, calcular la varianza de error a partir de la fórmula 5.3. Al resolver la fórmula 5.3 para s_{err} obtenemos:

$$s_{err} = s_{obs}\sqrt{1 - r_{11}} \qquad \textbf{(5.10)}$$

donde s es la desviación estándar de las calificaciones observadas de la prueba y r_{11} el coeficiente de confiabilidad test-retest. Este estadístico, conocido como *error estándar de medición* (s_{err}), es una estimación de la desviación estándar de una distribución normal de las calificaciones de

la prueba que se supone serían obtenidas por una persona que presentara la prueba un número infinito de veces. La media de esta distribución hipotética de calificaciones sería la calificación real de la persona en la prueba.

Para ilustrar el cálculo y el significado del error estándar de medición, suponga que la desviación estándar de una prueba es 6.63 y el coeficiente de confiabilidad test-retest es .85; entonces $s_{err} = 6.63\sqrt{1 - .85} = 2.57$. Si la calificación de una persona en la prueba es 40, puede concluirse, con 68% de confianza, que forma parte de un grupo de personas que tienen calificaciones observadas de 40 cuyas calificaciones reales en la prueba caen entre 40 – 2.57 = 37.43 y 40 + 2.57 = 42.57. Para obtener el intervalo de confianza de 95% para una calificación real, debe multiplicarse s_{err} por 1.96 y el producto resultante agregarse y restarse de la calificación observada: calificación observada ±1.96 s_{err}.

La figura 5.2 es un perfil o *psicógrafo* de las puntuaciones obtenidas por un estudiante de undécimo grado en las diez pruebas y tres compuestos de la Batería de Aptitudes Vocacionales de las Fuerzas Armadas (ASVAB). La puntuación del estudiante en una prueba o compuesto particular está indicada por las líneas verticales cortas que se proyectan a partir de la mitad de la barra horizontal correspondiente. La anchura de la barra horizontal es igual a 1.96 veces el error estándar de medición de esa prueba o compuesto particular. En consecuencia, podemos decir que hay una probabilidad de .95 de que la calificación real del estudiante en la prueba caiga den-

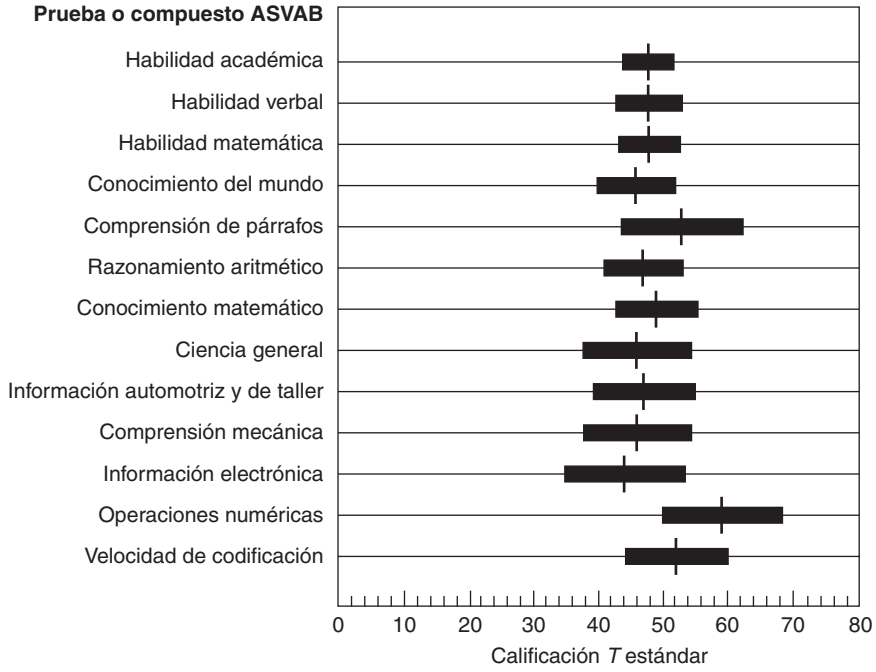


Figura 5.2 Gráfica de las calificaciones *T* de un estudiante (líneas verticales cortas que se proyectan desde la mitad de las barras horizontales) y barras que representan los intervalos de confianza del 95% para las calificaciones *T* reales del estudiante en las 10 pruebas y tres compuestos de la Batería de Aptitudes Vocacionales de las Fuerzas Armadas (ASVAB).

Vea el texto para detalles.

tro del rango numérico representado por la barra horizontal que se extiende desde la calificación observada $-1.96s_{err}$ hasta la calificación observada $+1.96s_{err}$.

Como regla empírica, la diferencia entre las puntuaciones de dos personas en la misma prueba no debe considerarse significativa salvo que sea por lo menos dos veces el error estándar de medición de la prueba. Por otro lado, la diferencia entre las puntuaciones de la misma persona en dos pruebas debe ser mayor que dos veces el error estándar de medición más grande para que la diferencia se interprete como significativa. Esto es así porque el error estándar de la diferencia entre las puntuaciones en las dos pruebas es mayor que el error estándar de medición de cada prueba.

Como vimos en la fórmula 5.10, el error estándar de medición se relaciona inversamente con el coeficiente de confiabilidad. Cuando $r_{11} = 1.00$, no hay error al estimar la calificación real de una persona a partir de su calificación observada; cuando $r_{11} = .00$, el error de medición alcanza su valor máximo (s). Por supuesto, una prueba que tiene un coeficiente de confiabilidad cercano a .00 es inútil porque la precisión de cualquier decisión tomada sobre la base de las puntuaciones estará al nivel del azar.

A diferencia de la teoría clásica de los tests, en la cual el error estándar de medición se aplica a todas las puntuaciones en una población particular, en la teoría de respuesta al ítem (IRT) difiere de una calificación a otra. En la IRT, el error estándar de medición de las puntuaciones correspondientes a un nivel particular de habilidad es igual al recíproco de la cantidad de información transmitida por una calificación a ese nivel. La cantidad de información proporcionada por las respuestas a un reactivo particular es determinada a partir de la función de información del reactivo, y la información proporcionada por la prueba como un todo en un nivel particular de habilidad es la suma de los valores de información del reactivo en ese nivel de habilidad (vea Hambleton, Swaminathan y Rogers, 1991).

Confiabilidad de las pruebas referidas a criterio

El concepto tradicional de confiabilidad corresponde a las pruebas referidas a normas, las cuales están diseñadas principalmente para diferenciar entre individuos que poseen varias cantidades de una característica específica. Entre mayor sea el rango de diferencias individuales en las puntuaciones de una prueba, mayor será la confiabilidad de la prueba. Por otro lado, al elaborar la mayoría de las pruebas referidas a criterio la meta es identificar a las personas como pertenecientes a uno de dos grupos. Un grupo consta de personas cuyas puntuaciones igualan o exceden el nivel de criterio (de dominio) en la habilidad que se está evaluando; el otro grupo consta de personas cuyas puntuaciones no alcanzan el nivel de criterio. En esta situación, resultan inapropiados los procedimientos correlacionales tradicionales para determinar los coeficientes test-retest, de formas paralelas y de consistencia interna.

El *coeficiente de acuerdo*, que es la proporción de calificaciones que caen por encima o por debajo del nivel de criterio en ambas aplicaciones o ambas formas, es un índice de la confiabilidad de una prueba referida a criterio. Otro índice es el *coeficiente kappa*, que es algo más difícil de calcular, pero estadísticamente más apropiado que el coeficiente de acuerdo (Cohen, 1968; Aiken, 1988).

Teoría de la generalización

Durante muchos años los psicómetras han enfatizado que una prueba no tiene una sino muchas confiabilidades, dependiendo de las varias fuentes de error de medición que se toman en consideración al calcular un coeficiente de confiabilidad. La muestra particular de reactivos incluidos

en la prueba, las instrucciones de aplicación, las condiciones ambientales (temperatura, iluminación, ruido) en que se aplica la prueba, y las idiosincracias y estados físicos o psicológicos temporales de los examinados pueden afectar la confiabilidad estimada de una prueba. Cualquiera de esas condiciones, o todas, puede contribuir a la varianza de error, simbolizada en la fórmula clásica 5.1 de la varianza de la prueba. Los procedimientos matemáticos de análisis factorial (vea el apéndice A) proporcionan una forma de dividir la varianza de la calificación real en la fórmula 5.1 en varianzas común y de factor específico, pero la fórmula no distingue entre varias fuentes de error que contribuyen a la varianza de error.

El hecho de que una prueba puede tener muchas confiabilidades, dependiendo de los efectos de varias fuentes de varianza de error, o *facetas*, ha sido incorporado en otros enfoques hacia la teoría de los tests . Uno de esos enfoques, la *teoría de la generalización*, empieza por conceptualizar una calificación observada de la prueba como una estimación de un *universo de calificaciones* correspondiente. El grado de precisión con el que la puntuación de la prueba estima el universo de puntuaciones depende de la naturaleza del universo, es decir, de las facetas particulares que lo definen.

Una puntuación de la prueba puede generalizarse a muchos universos diferentes, cada uno definido de acuerdo con cierta combinación de facetas. Las facetas que caracterizan a un universo dado, como las condiciones de aplicación de la prueba y la composición de reactivos o formas de ésta, pueden ser muy diferentes de las que definen a otros universos. Algunas de esas facetas pueden no tener efecto sobre la generalización de las puntuaciones de la prueba, mientras que los efectos de otras facetas pueden ser significativos.

Los cálculos de la teoría de la generalización implican aplicar las técnicas estadísticas de análisis de varianza para determinar la generalización de las puntuaciones de la prueba como función de los cambios en la(s) persona(s) que la presenta(n), diferentes muestras de reactivos que componen la prueba, las situaciones o condiciones en que se presenta ésta, y los métodos o personas involucrados en su calificación. Luego puede calcularse un *coeficiente de generalización*, el cual es similar al coeficiente tradicional de confiabilidad, como la razón de la varianza esperada de las calificaciones en el universo con la varianza de las calificaciones en la muestra. Por último, puede estimarse un *valor universal* de la calificación, similar a la calificación real de la teoría clásica de la confiabilidad (Cronbach, Gleser, Nanda y Rajaratnam, 1972).

Al enfatizar la importancia de las condiciones en que se administra una prueba y los propósitos para los que se diseñó, la teoría de la generalización ha cambiado el enfoque de los usuarios de las pruebas más allá de la preocupación con la prueba misma como un instrumento psicométrico bueno o malo en general a la pregunta de "¿Bueno o malo para qué propósito?"

La teoría de la generalización, la teoría de respuesta al ítem, el análisis de las estructuras de covarianza y otros métodos estadísticos modernos ciertamente tienen mayor complejidad técnica que la teoría clásica de los tests. Sea como sea, el desarrollo y las aplicaciones de una prueba todavía se basan en gran medida en los conceptos tradicionales de confiabilidad y validez y en los procedimientos derivados de ellos.

VALIDEZ

De manera tradicional, la *validez* se ha definido como el grado en que una prueba mide lo que está diseñada para medir. Una desventaja de esta definición es la implicación de que una prueba sólo tiene una validez, la cual supuestamente es establecida por un solo estudio para determinar si la prueba mide lo que se supone debe medir. En realidad, una prueba puede tener muchas cla-

ses de validez, dependiendo de los propósitos específicos para los cuales fue diseñada, la población objetivo, las condiciones en que se aplica y el método para determinar la validez.

Los métodos por los cuales puede determinarse la validez incluyen (1) analizar el contenido de la prueba, (2) calcular la correlación entre las calificaciones en la prueba y las calificaciones en el criterio de interés y (3) investigar las características psicológicas particulares o constructos medidos por la prueba. Todos esos procedimientos son útiles en la medida que mejoran la comprensión de lo que mide una prueba y proporcionan información para tomar decisiones sobre la gente. También puede ser de interés evaluar la *validez creciente* de una prueba, es decir, qué tanto añade la prueba a la predicción y comprensión de los criterios que ya son anticipados por otras medidas.

A diferencia de la confiabilidad, la cual es influida sólo por los errores no sistemáticos de medición, la validez de una prueba es afectada tanto por los errores no sistemáticos como por los sistemáticos (constantes). Por esta razón, una prueba puede ser confiable sin ser válida, pero no puede ser válida sin ser confiable. La confiabilidad es una condición necesaria, pero no suficiente, para la validez.

Validez de contenido

La apariencia física de una prueba con respecto a sus propósitos particulares (*validez de facie)* es ciertamente una consideración importante a tener en cuenta al comercializarla. No obstante, el concepto de *validez de contenido* se refiere a algo más que a la apariencia. La validez de contenido atañe a si éste produce un rango de respuestas que son representativas del dominio entero o universo de habilidades, entendimientos y otras conductas que supuestamente debe medir la prueba. Se supone que las respuestas a la muestra de reactivos de una prueba bien diseñada son indicativas de lo que serían las respuestas al universo entero de conductas de interés.

Un análisis de la validez de contenido ocurre más a menudo en conexión con las pruebas de aprovechamiento, para las cuales por lo general no existe un criterio externo especificado. La validez de contenido también es de interés en las medidas de aptitud, interés y personalidad, aunque quizá menos que la validez de constructo o la relacionada con un criterio. En una prueba de aprovechamiento se evalúa la validez de contenido analizando la composición de la prueba para determinar el grado en que representa los objetivos de la enseñanza. Una forma de lograr esto es comparar el contenido de la prueba con un bosquejo o tabla de especificaciones concernientes a la materia que va a ser cubierta por la prueba. Si expertos en la materia coinciden en que una prueba parece y actúa como un instrumento diseñado para medir lo que se supone debe medir, entonces se dice que posee validez de contenido. Dichos juicios involucran no sólo la apariencia de los reactivos de la prueba, sino también los procesos cognitivos implicados al responderlos. Es obvio que el proceso de evaluar la validez de contenido no debería esperar hasta que se haya elaborado la prueba. El juicio de los expertos en lo que se refiere a qué reactivos incluir es necesario desde el principio del proceso de elaboración de la prueba. Al definir el universo del contenido de la prueba y la muestra de ese universo que se va a incluir, los diseñadores establecen las condiciones indispensables para lograr construir un instrumento con validez de contenido.

Validez con relación a criterio

La validación de cualquier prueba de habilidad consiste en relacionar las calificaciones en la prueba con el desempeño en medidas o estándares de criterio con los cuales pueden compararse las calificaciones. Sin embargo, de manera tradicional, el término *validez con relación a criterio*

hace referencia a procedimientos en los cuales las calificaciones en la prueba de un grupo de personas se comparan con las puntuaciones, clasificaciones u otras medidas de desempeño. Algunos ejemplos de criterios contra los cuales se validan las pruebas son las notas escolares, calificaciones de un supervisor y número o cantidad de dólares en ventas. Siempre que se dispone de una medida de criterio en el momento de la prueba puede determinarse la *validez concurrente* del instrumento. Cuando las calificaciones en el criterio no están disponibles sino hasta cierto tiempo después de que se aplicó la prueba, se enfatiza la *validez predictiva* de la prueba.

Validez concurrente. Los procedimientos de validación concurrente se emplean siempre que una prueba se aplica a personas clasificadas en varias categorías, como grupos de diagnóstico clínico o niveles socioeconómicos, con el propósito de determinar si las puntuaciones en la prueba de las personas ubicadas en una categoría son significativamente diferentes de las de los individuos que se hallan en otras categorías. Si la calificación promedio varía de modo sustancial de una categoría a otra, entonces la prueba puede usarse como otra forma, quizá más eficiente, de asignar a la gente a esas categorías. Por ejemplo, las puntuaciones en el Inventario Multifásico de Personalidad de Minnesota (MMPI) han sido útiles en la identificación de trastornos mentales específicos, porque se ha encontrado que los pacientes a quienes los psiquiatras diagnostican trastornos particulares tienden a diferir del resto de la población en las puntuaciones que obtienen en ciertos grupos de reactivos (*escalas*).

Validez predictiva. La validez predictiva atañe a la precisión con que las puntuaciones de una prueba predicen puntuaciones de criterio, según lo indica la correlación entre la prueba (predictor) y un criterio del desempeño futuro. La validez predictiva es de interés sobre todo para las pruebas de aptitud o de inteligencia, ya que las puntuaciones en esos tipos de instrumentos a menudo se correlacionan con las puntuaciones, notas de cursos, calificaciones de pruebas de aprovechamiento y otros criterios de desempeño.

La magnitud de un coeficiente de validez predictiva está limitada por la confiabilidad de las variables de predicción y de criterio; no puede ser mayor que la raíz cuadrada del producto de esas dos confiabilidades. Por ésta y por otras razones, la correlación entre un predictor y una variable de criterio, calculada mediante procedimientos descritos en el apéndice A, varía con el criterio específico, pero rara vez es mayor a .60. Como la proporción de la varianza en la variable de criterio que puede ser explicada por la variación en la variable predictora es igual al cuadrado de la correlación entre las variables predictora y de criterio, por lo general no puede predecirse más de 36% de la variación en las puntuaciones de criterio a partir de las puntuaciones obtenidas en una prueba u otro instrumento psicométrico. Esto deja sin explicar o predecir 64% de la varianza de criterio. Considerando que la validez predictiva de la mayoría de las pruebas es menor de .60, es comprensible por qué deben hacerse con cuidado las afirmaciones concernientes a la posibilidad de predecir los criterios de desempeño a partir de las puntuaciones obtenidas en las pruebas psicológicas.

Error estándar de estimación. La sección sobre regresión y predicción en el apéndice A describe el procedimiento a seguir para determinar una ecuación de regresión (ecuación de predicción) y pronosticar las calificaciones de criterio de un grupo de personas a partir de sus puntuaciones en pruebas o en otras variables. Sin embargo, ingresar la puntuación de una persona en una prueba a una ecuación de regresión sólo produce un estimado de la calificación que la persona obtendrá en realidad en la variable de criterio. Si la calificación de criterio que se predice para una persona se considera como la media de una distribución normal de las calificacio-

nes de criterio obtenidas por un grupo de individuos que obtuvieron la misma calificación que la persona en la variable predictora, entonces la desviación estándar de esta distribución es un índice del error promedio en esas predicciones. Este estadístico, conocido como *error estándar de estimación* (s_{est}), es aproximadamente igual a:

$$s_{est} = s\sqrt{1 - r^2} \quad \textbf{(5.11)}$$

donde s es la desviación estándar de la calificación de criterio y r es la correlación producto-momento entre el predictor (prueba) y el criterio.

Por ejemplo, suponga que la desviación estándar de cierta medida de criterio es 15 y la correlación entre las puntuaciones de la prueba y de criterio es .50; entonces $s_{est} = 15\sqrt{1 - 50^2} = 13$. Si la calificación de criterio que se predice de un sujeto es 50, hay una posibilidad de 68 entre 100 de que la persona obtendrá una calificación de criterio entre 37 y 63 ($Y_{pred} \pm s_{est}$), y aproximadamente 95 de 100 de que obtendrá una calificación de criterio entre 25 y 75 ($Y_{pred} \pm 1.96\ s_{est}$). De manera más precisa, las posibilidades son de 68 entre 100 de que la persona forme parte de un grupo de individuos que tienen una calificación de criterio pronosticada de 50 cuyas calificaciones de criterio obtenidas cayeron entre 37 y 63. De manera similar, hay una posibilidad aproximada de 95 entre 100 de que el individuo forme parte de un grupo de personas con una calificación promedio anticipada de 50 cuyas calificaciones de criterio obtenidas cayeron entre 25 y 75. Como lo ilustra este ejemplo, cuando la correlación entre las calificaciones de la prueba y de criterio es baja, la calificación de criterio obtenida por una persona puede ser muy diferente de la pronosticada. Por esta razón, debe tenerse cautela al interpretar las calificaciones predichas cuando la correlación entre la prueba y la medida de criterio es modesta. Entre menor sea el coeficiente de correlación, más grande es el error estándar de estimación y menos precisa es la predicción de la prueba al criterio.

Factores que afectan la validez con relación a criterios

La validez con relación a criterios de una prueba puede ser influida por una serie de factores, incluyendo las diferencias de grupo, la extensión de la prueba, la contaminación del criterio y la tasa base. La validez creciente de una prueba, es decir, la contribución de la prueba que excede a las contribuciones de otras variables, también debería ser considerada al decidir si se va a utilizar la prueba con propósitos de selección y ubicación.

Diferencias de grupo. Las características de un grupo de personas en quienes se valida una prueba incluyen variables como sexo, edad y rasgos de personalidad. Esos factores, que en este contexto se conocen como *variables moderadoras*, pueden afectar la correlación entre una prueba y una medida de criterio. La magnitud de un coeficiente de validez, como la de un coeficiente de confiabilidad, también está influida por el grado de heterogeneidad del grupo de validación en lo que mide la prueba. Los coeficientes de validez tienden a ser más pequeños en los grupos más homogéneos, es decir, los grupos que tienen un rango más estrecho de calificaciones. El tamaño de un coeficiente de correlación es una función de las variables de predicción y de criterio, por lo que estrechar el rango de calificaciones en cualquier variable tiende a disminuir el coeficiente de validez predictiva.

Como la magnitud de un coeficiente de validez varía con la naturaleza del grupo probado, una prueba recién elaborada que resulte ser un predictor válido de una variable de criterio particular en un grupo de gente debe tener una validación cruzada en un segundo grupo. En la *valida-*

ción cruzada se aplica una prueba a una segunda muestra de personas para determinar si conserva su validez entre muestras diferentes. Debido a la operación de los factores del azar, la magnitud de un coeficiente de validez por lo general se reduce en alguna medida en la validación cruzada. En consecuencia, en la mayor parte de los casos se considera que la correlación entre las calificaciones de predicción y de criterio en la validación cruzada es un mejor indicador de la validez predictiva que la correlación original prueba-criterio. La validación cruzada, que es una manera de determinar la *generalización de la validez* de una prueba, es decir, si la prueba sigue siendo válida en situaciones diferentes, también puede involucrar una muestra diferente (paralela) de reactivos. Con diferentes muestras de examinados, diferentes muestras de reactivos de la prueba, o en ambos casos, suele darse alguna reducción del coeficiente de validez en la validación cruzada. Se han propuesto fórmulas para "corregir" dicha reducción, pero implican ciertas suposiciones que no siempre se cumplen.

Extensión de la prueba. Al igual que la confiabilidad, la validez varía directamente con la extensión de la prueba y con la heterogeneidad del grupo de personas examinadas. Hasta cierto punto, las puntuaciones en una prueba más larga y en una prueba administrada a un grupo de individuos que varían de manera considerable en las características a medir tienen varianzas más grandes y, en consecuencia, mayor validez predictiva que las puntuaciones de pruebas más cortas o de pruebas aplicadas a grupos más homogéneos. Se han propuesto fórmulas que corrigen los efectos que tienen en la validez la restricción de rangos de calificación y la extensión acortada de la prueba, pero sólo son apropiadas bajo ciertas circunstancias especiales.

Contaminación de criterios. La validez de una prueba está limitada no sólo por su confiabilidad y el criterio, sino también por la validez del propio criterio como medida de la variable de interés. En ocasiones el criterio se hace menos válido, o se *contamina*, por el método particular de medir las calificaciones de criterio. Por ejemplo, un psicólogo clínico enterado de que un grupo de pacientes ya ha sido diagnosticado como psicótico puede percibir mal los signos psicóticos en las respuestas de esos pacientes a las pruebas de personalidad. Entonces el *método de comparación de grupos,* en el cual se comparan las calificaciones obtenidas por los psicóticos en la prueba con las obtenidas por los normales, arrojará evidencia falsa a favor de la validez de la prueba. Dicha contaminación del criterio (psicóticos contra normales) puede controlarse por medio de un *análisis ciego*, esto es, haciendo que quien emite el diagnóstico no disponga de información acerca de los examinados a excepción de sus puntuaciones en la prueba. Sin embargo, muchos psicólogos clínicos sostienen que el análisis ciego no es natural ya que no es la forma en que las pruebas se emplean en realidad en los escenarios clínicos.

Validez creciente. Cuando se intenta decidir si la aplicación de un instrumento particular de evaluación con propósitos predictivos o de diagnóstico está justificada por su costo, también debería considerarse la *validez creciente*. La validez creciente se refiere a la cuestión de qué tanta precisión más tienen las predicciones y los diagnósticos cuando se incluye una prueba particular en una batería de procedimientos de evaluación. Es posible que otros métodos de evaluación menos costosos (observación, entrevista, inventario biográfico) puedan satisfacer los propósitos de la evaluación igual de bien sin usar una prueba adicional. La validez creciente se relaciona con el concepto de utilidad, tal como se aplica en los contextos de selección de personal. La *utilidad* de una prueba se define como un incremento medido en la calidad de los empleados que son contratados o promovidos sobre la calidad de los empleados cuando no se usa una prueba u otro procedimiento de evaluación (Cascio, 2000).

Validez de constructo

La validez predictiva es del mayor interés en la selección y ubicación en un contexto ocupacional o educativo. Diferentes tipos de pruebas de habilidad, y en ocasiones pruebas de personalidad y de interés, se utilizan con propósitos de selección y ubicación. La validez de constructo es de un interés mayor aún con respecto a las pruebas de personalidad. La *validez de constructo* de un instrumento de evaluación psicológica se refiere al grado en que el instrumento mide un *constructo* particular, o concepto psicológico como la ansiedad, la motivación para el logro, la extroversión-introversión o el neuroticismo. La validez de constructo, que es el tipo más general de validez, no se determina de una sola manera o por una investigación. Más bien involucra una red de investigaciones y otros procedimientos diseñados para determinar si un instrumento de evaluación que supuestamente mide una determinada variable de personalidad en realidad lo hace.

Evidencia a favor de la validez de constructo. Entre las fuentes de evidencia a favor de la validez de constructo de una prueba se encuentran las siguientes:

1. Los juicios de expertos de que el contenido de la prueba corresponde al constructo de interés.
2. Análisis de la consistencia interna de la prueba.
3. Estudios, tanto en grupos formados de manera experimental como en grupos que se presentan de manera natural, de las relaciones entre las puntuaciones de la prueba y otras variables en las cuales difieren los grupos.
4. Correlaciones de las puntuaciones en la prueba con las puntuaciones en otras pruebas y variables con las cuales se espera que tengan cierta relación, seguidas por un análisis factorial de esas correlaciones.
5. Interrogar con detalle a los examinados o a los calificadores acerca de sus respuestas a una prueba o escala de calificación para revelar los procesos mentales específicos implicados al dar respuesta a los reactivos.

Como lo revela esta lista, varios tipos de información contribuyen al establecimiento de la validez de constructo de un instrumento psicométrico. La información puede obtenerse de análisis racionales o estadísticos de las variables evaluadas por el instrumento y por estudios de su capacidad para predecir la conducta en las situaciones en que opera el constructo.

Las demostraciones experimentales como las usadas en la validación de constructo de la Escala de Taylor de Ansiedad Manifiesta (TMAS) (Taylor, 1953) son particularmente importantes en el establecimiento de la validez de constructo. De acuerdo con la teoría hulliana del aprendizaje, la ansiedad es una pulsión y, por consiguiente, la gente muy ansiosa debe condicionarse con mayor facilidad que la gente menos ansiosa. Suponiendo que esta teoría es correcta, los individuos que tienen un alto nivel de ansiedad deben adquirir —con más rapidez que quienes tienen un bajo nivel de ansiedad— un parpadeo condicionado en una situación de condicionamiento clásico donde estén presentes una luz, un soplo de aire y el parpadeo. Por lo tanto, si es una medida válida del constructo de *ansiedad*, quienes obtienen puntuaciones altas en la escala TMAS deberían condicionarse con mayor facilidad en esta situación que quienes obtienen bajas puntuaciones. La verificación de esta predicción contribuyó de manera significativa a aceptar la validez de constructo de la TMAS.

Validación convergente y discriminante. Un instrumento con validez de constructo debe tener correlaciones altas con otras medidas o métodos de medición del mismo constructo (*validez convergente*), pero correlaciones bajas con las medidas de constructos diferentes (*validez discri-*

minante). La evidencia a favor de estas validaciones de un instrumento psicométrico puede obtenerse comparando las correlaciones entre las medidas de:

1. El mismo constructo usando el mismo método.
2. Diferentes constructos usando el mismo método.
3. El mismo constructo usando métodos diferentes.
4. Diferentes constructos usando métodos diferentes.

La validez de constructo de un instrumento psicométrico se confirma por este *planteamiento de características y métodos múltiples* (Campbell y Fiske, 1959) cuando las correlaciones entre el mismo constructo medidas por el mismo y por diferentes métodos son significativamente mayores que las correlaciones entre diferentes constructos medidas por los mismos o por diferentes métodos. Por desgracia, los resultados de dichas comparaciones no siempre resultan de esta manera. Ocasionalmente las correlaciones entre diferentes constructos medidas por el mismo método son más altas que las correlaciones entre el mismo constructo medidas por métodos diferentes. Esto significa que el método (inventario de lápiz y papel, técnica proyectiva, escala de calificación, entrevista, etc.) es más importante que el constructo o rasgo particular en la determinación de lo que está siendo medido que el constructo o rasgo que supuestamente está siendo evaluado.

UTILIZACIÓN DE TESTS EN LA TOMA DE DECISIONES DEL PERSONAL

Desde la antigüedad las personas han sido seleccionadas, clasificadas y ubicadas en determinados puestos para realizar varias tareas. Sin embargo, con frecuencia los procedimientos seguidos para seleccionar, clasificar y ubicar personal han sido azarosos y asistemáticos. Se ha empleado gran variedad de procedimientos para la selección y valoración de personal, muchos de los cuales se basan en la observación casual y la intuición. Por ejemplo, en un tiempo se asignó gran importancia a rasgos físicos como la forma de la cabeza, los movimientos oculares y la apariencia corporal general. El origen étnico, la posición social y las conexiones sociales también influyeron en la determinación de quién era designado para ocupar cierto puesto, contratado para un trabajo específico o aceptado en determinado programa educativo.

Detección

De manera tradicional, la selección de personal se ha interesado en identificar, de entre un grupo de solicitantes, a los que son más capaces de realizar las tareas designadas. En este enfoque se utilizan las pruebas psicológicas, junto con información que no proviene de la prueba (historia personal, características físicas, recomendaciones, etc.), para ayudar a seleccionar a los solicitantes que pueden desempeñar trabajos particulares, ya sea de manera inmediata o luego de un entrenamiento apropiado.

Un procedimiento de selección de personal puede ser bastante simple o muy complejo, dependiendo de la naturaleza de la organización y de la tarea para la cual se están seleccionando los solicitantes. El planteamiento más directo es la estrategia de hundirse o nadar en la cual todos los solicitantes son seleccionados o admitidos, pero sólo se conserva a quienes tienen un desempeño efectivo. En algunas formas ésta es una estrategia ideal de selección, pero también es costosa tanto para la organización como para los solicitantes. En consecuencia, casi todas las organizaciones grandes utilizan actualmente algún tipo de procedimiento de *detección* por el cual los solicitantes que son

claramente inadecuados para la tarea (trabajo, programa, etc.) son rechazados de inmediato. Si el instrumento de detección es un instrumento psicométrico de algún tipo, se acepta a los solicitantes que obtienen una calificación mínima especificada (*calificación límite*) o más alta en la prueba, mientras que se rechaza a los que puntúan por debajo de la calificación límite. Este procedimiento es bastante impersonal, y en ocasiones puede parecer duro desde la perspectiva de los solicitantes. Pero las organizaciones funcionan de manera más eficiente cuando los empleados poseen las habilidades indispensables para realizar de manera efectiva las tareas asignadas.

Clasificación y ubicación

La detección inicial, por lo regular, es seguida por la *clasificación* y la asignación de los solicitantes seleccionados a una de varias categorías ocupacionales. Las decisiones de clasificación pueden implicar el agrupamiento de los empleados sobre la base de sus puntuaciones en más de una prueba psicológica, como la asignación de los reclutas militares a especialidades ocupacionales de acuerdo con sus calificaciones en la Batería de Aptitudes Vocacionales de las Fuerzas Armadas. La detección y la clasificación con frecuencia son seguidas por la *ubicación* de los seleccionados en un nivel particular de determinado trabajo o programa.

El proceso de selección de personal consiste usualmente en una secuencia de etapas vinculadas a una serie de decisiones de sí-no (pase-fracaso) basadas en la información obtenida de formularios de solicitud, cartas de referencia, llamadas telefónicas, entrevistas personales, observaciones y pruebas psicológicas. El propósito de recabar dicha información es idéntico al de cualquier otra aplicación de la psicología: hacer mejores predicciones de la conducta futura sobre la base de la conducta pasada y presente. Entre más confiable y válida sea la información, mayor es la probabilidad de hacer predicciones precisas de la conducta en el trabajo o en el programa y, por ende, resultan más adecuadas las decisiones de selección. Por supuesto, la confiabilidad y validez de los instrumentos y procedimientos de evaluación psicológica para tomar decisiones de selección no pueden determinarse sólo mediante la inspección de los materiales de evaluación. La confiabilidad y la validez deben evaluarse de manera empírica, lo cual es una de las tareas propias de los psicólogos organizacionales.

Una tabla de expectativas

Cuando se utilizan las pruebas con propósitos de selección no es esencial determinar la correlación prueba-criterio ni la ecuación de regresión que vincula el desempeño en la variable de criterio con las calificaciones en la prueba. Los métodos correlacionales pueden aplicarse a la elaboración de tablas de expectativas teóricas, pero es posible elaborar una tabla de expectativas empíricas sin calcular un coeficiente de correlación o cualquier otro estadístico a excepción de frecuencias y porcentajes. Suponga, por ejemplo, que la tabla 5.1 fue elaborada a partir de una distribución conjunta de frecuencia de las calificaciones de 250 solicitantes de empleo en una Prueba de Selección Ocupacional (OST), y de las puntuaciones -asignadas a los solicitantes por sus supervisores laborales seis meses después de haber sido contratados. Los intervalos de calificación de la OST se presentan al lado izquierdo de la tabla y las puntuaciones de desempeño (en una escala de 1 a 8) a lo largo de la parte superior. Las frecuencias sin cursivas colocadas en las celdas de la tabla representan la cantidad de empleados que obtuvieron tanto puntuaciones en la OST, dentro de un rango especificado de 5 puntos, como las puntuaciones de desempeño indicadas en la parte superior de la columna. Por ejemplo, 10 empleados cuyas puntuaciones en la OST estuvieron entre 81 y 85 recibieron de sus supervisores una calificación de desempeño de 5, mientras que 14 empleados cuyas puntuaciones en la OST cayeron entre 66 y 70 recibieron una calificación de desempeño de 4.

TABLA 5.1 Tabla empírica de expectativas

	CALIFICACIÓN DEL DESEMPEÑO							
CALIFICACIÓN EN LA PRUEBA DE SELECCIÓN OCUPACIONAL	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
96–100						(*100*)/1		(*67*)/2
91–95					(*100*)/2		(*82*)/5	(*36*)/4
86–90				(*100*)/1	(*94*)/8	(*50*)/3	(*33*)/4	(*11*)/2
81–85				(*100*)/4	(*85*)/10	(*48*)/7	(*22*)/5	(*4*)/1
76–80			(*100*)/6	(*88*)/12	(*63*)/16	(*31*)/13	(*4*)/2	
71–75		(*100*)/4	(*94*)/7	(*83*)/25	(*45*)/21	(*12*)/5	(*5*)/3	
66–70		(*100*)/5	(*87*)/10	(*61*)/14	(*24*)/7	(*5*)/2		
61–65	(*100*)/1	(*96*)/6	(*72*)/8	(*40*)/5	(*20*)/4	(*4*)/1		
56–60	(*100*)/2	(*85*)/5	(*46*)/4	(*15*)/2				
51–55	(*100*)/1							

Los números en cursivas y entre paréntesis de la tabla 5.1 son los porcentajes de la gente con puntuaciones OST, en un intervalo determinado, cuyas puntuaciones de desempeño fueron iguales o mayores al valor correspondiente en las celdas dadas. De este modo, 85% de los empleados cuyas puntuaciones en la OST cayeron en el intervalo 81 a 85 recibieron de desempeño de 5 o más altas, y 61% de los que obtuvieron puntuaciones en la OST entre 66 y 70 tuvieron calificaciones de desempeño de 4 o más altas.

Para ilustrar cómo se aplica este tipo de información al proceso de selección ocupacional, suponga que Juan, un empleado potencial de un grupo similar al grupo para el cual se elaboró la tabla 5.1, obtiene una puntuación de 68 en la Prueba de Selección Ocupacional. Entonces puede estimarse que Juan tiene una posibilidad de 61 en 100 de recibir de su supervisor una calificación de 4 o más alta en el desempeño en el trabajo seis meses después de empezar éste, pero sus posibilidades de obtener una calificación del desempeño de 6 o más alta son sólo de 5 en 100. Si una calificación de 4 o más alta es aceptable, es probable que Juan sea contratado.

Factores que afectan la precisión predictiva

La precisión con la que puede predecirse la calificación de criterio de un solicitante no sólo depende del tamaño de la correlación entre las variables de predicción y de criterio, sino también del número de errores por falsos-positivos y falsos-negativos, la razón de selección, y la tasa base. Si en una prueba se establece una calificación límite muy baja, habrá muchas aceptaciones incorrectas o *falsos positivos*; esto es, solicitantes que fueron seleccionados pero que no tuvieron éxito en el trabajo o en el programa. Por otro lado, si se establece una calificación límite muy alta, habrá muchos rechazos incorrectos o *falsos negativos*; esto es, solicitantes que no fueron seleccionados pero que de haberlo sido habrían tenido éxito. Como el propósito de la selección de personal es obtener tantos "aciertos" como sea posible (rechazar a los fracasos potenciales y seleccionar a los éxitos potenciales), la calificación límite debe establecerse con cuidado.

Para ilustrar estos conceptos, vaya de nuevo a la tabla 5.1. Suponga que la calificación límite en la OST se establece en 66 y que 4 se considera una calificación mínima aceptable de desempeño en el trabajo. Entonces $4 + 5 + 6 + 7 + 10 = 32$ de los empleados representados en la tabla 5.1 serán clasificados como falsos positivos: tuvieron una calificación de al menos 66 en la OST, pero tuvieron calificaciones de desempeño de menos de 4. Por otro lado, $5 + 2 + 4 + 1 = 12$ empleados serán falsos negativos: calificaron por debajo de 66 en la OST, pero recibieron calificaciones de desempeño de 4 o más altas. Observe que al elevar la calificación límite en la OST disminuye el número de falsos positivos, pero incrementa el número de falsos negativos. El efecto opuesto, un incremento en los falsos positivos y una disminución en los falsos negativos, ocurre cuando se baja la calificación límite en la OST.

Otro factor importante a considerar al establecer la calificación límite en una prueba o prueba compuesta es la *razón de selección*, que es la proporción de solicitantes que serán seleccionados. Entre menor sea la razón de selección, más alta es la calificación límite y viceversa. Como el número de errores por falsos positivos y falsos negativos es afectado dependiendo de dónde se establezca la calificación límite, podemos argumentar que la razón de selección debería ser determinada por la gravedad relativa de esos dos tipos de error. ¿Es el error cometido al aceptar a un solicitante que no logra realizar el trabajo de manera satisfactoria (falso positivo) más o menos grave que rechazar a un solicitante que podría haberlo realizado con éxito si hubiera sido seleccionado (falso negativo)? Dichos errores deberían ser tomados en cuenta, pero el número total de solicitantes es al menos igual de importante al determinar la razón de selección. Por ejemplo, cuando el mercado de trabajo es cerrado, el número de solicitantes será pequeño. Entonces la razón de selección necesitará ser alta y, en consecuencia, la calificación límite en la prueba debe ser lo bastante baja como para obtener el número deseado de personas. Por otro lado, en un mercado laboral libre o abierto, el número de solicitantes es grande, por lo que la razón de selección será baja. Una razón de selección baja significa que será necesario establecer una calificación límite bastante alta en la prueba, lo que dará lugar a un número menor de solicitantes aceptados y falsos positivos y a un número mayor de solicitantes rechazados y falsos negativos. El porcentaje de solicitantes exitosos varía inversamente con la razón de selección, pero varía directamente con la validez de la prueba u otros instrumentos de selección. En general, una prueba más válida conduce a un porcentaje más grande de aciertos y a un porcentaje más pequeño de falsos positivos y falsos negativos.

Un factor más que también afecta la precisión con que una prueba puede identificar a las personas que se comportarán de cierta manera es la *tasa base*, esto es, la proporción de solicitantes que se esperaría desempeñaran satisfactoriamente un trabajo incluso si no se hubiera empleado un instrumento o procedimiento de selección. Como con la razón de selección, una prueba diseñada para

predecir un tipo particular de conducta es más efectiva cuando la tasa base es 50% y menos efectiva cuando la tasa base es muy alta o muy baja. Por ello, una prueba diseñada con el propósito de seleccionar gente para un trabajo muy complejo, en el cual relativamente pocos solicitantes pueden tener un buen desempeño, no sería tan efectiva como una diseñada para seleccionar gente para un trabajo en el cual la mitad de la población de solicitantes puede tener un desempeño satisfactorio. El concepto de tasa base no se limita a la selección de personal; también es importante en el diagnóstico clínico. Por ejemplo, debido a que la incidencia de suicidios en la población general es muy baja, una prueba diseñada para identificar a personas suicidas no sería muy exacta. Se esperaría un mejor resultado de una prueba diseñada para identificar a neuróticos porque el porcentaje de neuróticos en la población general es más alto que el de suicidas potenciales.

La cantidad de información aportada por una prueba más allá de la tasa base puede ser determinada consultando la *tabla Taylor-Russell* para la tasa base especificada (Taylor y Russell, 1939). La tabla presenta el porcentaje de solicitantes seleccionados que puede esperarse tengan éxito en un trabajo, o en otra situación de selección, como función del coeficiente de validez de la prueba, la tasa base y la razón de selección. La inspección de varias tablas Taylor-Russell para tasas base específicas muestra que el porcentaje de solicitantes que se espera tengan éxito varía directamente con el coeficiente de validez, pero inversamente con la razón de selección. En general, en una tasa base intermedia y con una razón de selección baja, las calificaciones en una prueba que tiene un coeficiente de validez modesto pueden producir un incremento sustancial en el número de aciertos en una situación de selección.

El uso de las tablas Taylor-Russell supone una definición clara, discreta y dicotómica del éxito (contra el fracaso) en una situación de selección. Se han elaborado enfoques similares que implican criterios continuos de éxito basados en la teoría de la decisión y la utilidad, pero son complejos y escapan al alcance de este libro (vea Cascio y Ramos, 1986; Cronbach y Gleser, 1965; Raju, Normand y Burke, 1990; Schmitt y Robertson, 1990).

Límite múltiple y regresión múltiple

Establecer la calificación límite para una prueba de selección o ubicación es un proceso complejo de juicio. Además de los factores analizados líneas arriba, la calificación límite y la utilidad de una prueba en general son afectadas por otros tipos de información del solicitante.

Con frecuencia, un conjunto de calificaciones de prueba y otras medidas se combinan para tomar decisiones de selección y clasificación. Un procedimiento para combinar calificaciones, conocido como *obstáculos sucesivos* o *límites múltiples*, establece calificaciones límite separadas en cada una de varias medidas. Entonces un solicitante debe puntuar en el punto límite o por arriba de éste en cada medida separada en una situación donde una alta calificación en una medida no compensa una baja calificación en otra medida. Por ejemplo, la habilidad para diferenciar entre tonos de diferentes alturas es esencial para el desempeño efectivo de un director de orquesta. Independientemente de qué tan altas puedan ser sus calificaciones en pruebas de habilidades cognoscitivas, no puede esperarse que las personas sordas a los tonos sean buenos directores de orquesta.[1]

[1]Puede desafiar la imaginación, pero han existido algunas excepciones notables al requisito de que los ejecutantes musicales deberían tener buena audición. En música, como en otras actividades profesionales, la gente puede ser capaz de compensar las discapacidades sensoriales o motrices enfatizando otras capacidades que permanecen intactas. Además, al igual que Demóstenes, pueden sobrecompensar. Se dice que Demóstenes superó un defecto del habla al colocar guijarros en su boca y rugir a las olas, convirtiéndose en uno de los más grandes oradores de la antigua Grecia.

Un enfoque más matemático de la combinación de las calificaciones de una muestra grande de personas en varias medidas es determinar una *ecuación de regresión múltiple* en la cual se apliquen diferentes pesos asignados estadísticamente a las calificaciones en diferentes pruebas. Una vez que se han determinado los pesos de regresión, puede calcularse para cada solicitante una sola calificación de criterio pronosticada multiplicando la calificación del solicitante en cada variable por el peso apropiado, sumando los productos y restando una constante. Por ejemplo, una ecuación de regresión múltiple empleada para propósitos de admisión en una universidad fue el GPA_{pred} = .002(SAT-V) + .001(SAT-M) + .030(HSR) - 2.00, donde SAT-V y SAT-M son las calificaciones del solicitante en las secciones Verbal y Matemática de la Prueba de Evaluación Escolar, HSR es una calificación *T* del rango del solicitante en su clase de graduación de la preparatoria, y GPA_{pred} es el promedio académico pronosticado en el primer año del solicitante en la universidad. Si las calificaciones de un solicitante en particular en las dos secciones del SAT son 600 y 500 y su rango en la preparatoria es 70, entonces su promedio académico pronosticado es GPA_{pred} = .002(600) + .001(500) + .030(70) –2.00 = 1.8, lo cual equivale a una C baja.

En el enfoque de regresión múltiple, una calificación alta en una variable predictora puede compensar una calificación baja en otra variable predictora. En consecuencia, este planteamiento no debería usarse cuando una calificación mínima en cualquiera de los predictores sea esencial para el desempeño efectivo en el criterio. Cuando se utiliza un enfoque de regresión múltiple, debe calcularse un *coeficiente de correlación múltiple (R)*, el cual es un indicador de la relación de una combinación ponderada de las variables predictoras con la variable de criterio.

RESUMEN

La confiabilidad se refiere a la libertad relativa que tienen las calificaciones de prueba de los errores de medición. En la teoría clásica de las calificaciones de prueba, la confiabilidad se define como la razón de la varianza de la calificación real en una prueba con la varianza de su calificación observada. Dado que la varianza de la calificación real no puede calcularse directamente, la confiabilidad debe estimarse mediante uno de varios procedimientos que toman en consideración varias fuentes de error de medición. Tres métodos tradicionales para estimar la confiabilidad de una prueba u otro instrumento de evaluación son test-retest, formas paralelas y consistencia interna. El método de formas paralelas, que tiene en consideración los errores debidos a diferentes momentos de aplicación, así como los debidos a diferentes muestras de los reactivos de prueba, es el más satisfactorio. Debido a que la elaboración de las formas paralelas es costosa y consume tiempo, los procedimientos de test-retest y de consistencia interna son las fuentes más populares de evidencia de confiabilidad. Los enfoques de consistencia interna, que son menos apropiados para las pruebas de velocidad, comprenden la división por mitades, las fórmulas Kuder-Richardson y el coeficiente alfa.

El error estándar de medición, que varía inversamente con la magnitud del coeficiente de confiabilidad, se emplea al calcular intervalos de confianza para las calificaciones reales en una prueba. Entre más grande sea el error estándar de medición, más amplio es el rango de calificaciones que puede decirse, con un grado especificado de confianza, contiene la calificación real de un examinado en la prueba.

La confiabilidad de una prueba varía directamente con el número de reactivos y la heterogeneidad del grupo que la presenta. La confiabilidad también varía con el nivel de dificultad de los reactivos que componen la prueba, siendo más alta con reactivos de dificultad intermedia.

En este capíulo se analizaron de manera breve los procedimientos para determinar la consistencia entre diferentes calificadores (confiabilidad entre calificadores) y la confiabilidad de las pruebas referidas a criterio. También se prestó atención a la teoría de la generalización, la cual conceptualiza la calificación de una prueba como la muestra de una población y, por ende, como el estimado de una calificación real o valor universal.

La confiabilidad es una condición necesaria pero no suficiente para lograr la validez, que es el grado en el cual una prueba mide lo que está diseñada para medir. La información sobre la validez de una prueba puede obtenerse de varias maneras: analizando el contenido de la prueba (validez de contenido), correlacionando las calificaciones de la prueba con calificaciones en un criterio medidas al mismo tiempo (validez concurrente), correlacionando las calificaciones de la prueba con calificaciones en un criterio medidas en un momento posterior (validez predictiva), y por el estudio sistemático de lo adecuado de la prueba para valorar un constructo psicológico especificado (validez de constructo). En las pruebas de aprovechamiento, por lo regular, se valida el contenido, mientras que la validez predictiva es de mayor interés con respecto a las pruebas de aptitud. La validez concurrente y la de constructo son importantes para las pruebas de personalidad.

La magnitud de un error cometido al predecir la calificación de criterio de una persona a partir de su calificación en una prueba es calculada mediante el error estándar de estimación, el cual varía inversamente con el tamaño del coeficiente de validez relacionado con el criterio. Tanto el coeficiente de validez relacionado con el criterio como el error estándar de estimación son afectados por varios factores que comprenden las diferencias de grupo, la extensión de la prueba y la contaminación del criterio. Como la magnitud de un coeficiente de validez puede ser afectada por factores aleatorios, las pruebas usadas con propósitos predictivos deberían someterse a validación cruzada en muestras separadas de personas. También es importante considerar cuánto contribuyen las calificaciones de prueba al proceso de tomar buenas decisiones acerca de la gente mucho más allá de las contribuciones de otras variables (validez creciente).

La información sobre la validez de constructo de una prueba como medida de una variable o característica psicológica particular puede obtenerse de varias maneras. En particular, es útil un análisis de correlación entre la prueba y otras medidas del mismo constructo obtenidas por el mismo método o por métodos diferentes, así como medidas de diferentes constructos obtenidas por el mismo método o por métodos diferentes (matriz de rasgos y métodos múltiples).

Las pruebas psicológicas se aplican en escenarios ocupacionales con propósitos de selección, clasificación, promoción y valoración periódica de empleados. Algunos de los procedimientos estadísticos que se utilizan con esos propósitos son tablas de expectativas, razones de selección y métodos de límites múltiples y regresión múltiple.

PREGUNTAS Y ACTIVIDADES

1. Calcule los coeficientes de confiabilidad de división por mitades (nones y pares) y Kuder-Richardson (fórmulas 20 y 21) en las siguientes calificaciones de diez examinados a diez reactivos en una prueba de aprovechamiento donde 1 indica una respuesta correcta y 0 una respuesta errónea.

	EXAMINADO									
REACTIVO	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
1	1	1	0	1	1	0	1	0	1	0
2	1	0	0	0	0	1	0	0	0	1
3	1	1	1	1	1	0	1	0	0	0
4	1	1	1	0	0	1	0	1	0	0
5	1	0	1	1	0	0	0	0	0	0
6	1	1	1	0	1	1	1	0	0	0
7	1	0	1	1	0	0	1	1	0	1
8	1	1	1	0	1	1	0	0	1	0
9	1	1	0	1	1	1	0	1	0	0
10	1	1	1	1	1	0	0	0	1	0
Totales	10	7	7	6	6	5	4	3	3	2

La media ($\overline{X}$) de las calificaciones totales es 5.30 y la varianza (s^2) es 5.21.

2. Calcule el error estándar de medición (s_{err}) de una prueba que tiene una desviación estándar de 10 y un coeficiente de confiabilidad de formas paralelas de .84. Luego use el valor obtenido de s_{err} para encontrar el intervalo de confianza de 95% para las calificaciones reales correspondientes a las calificaciones obtenidas de 40, 50 y 60.

3. Una prueba que consta de 40 reactivos tiene un coeficiente de confiabilidad de .80. ¿Aproximadamente cuántos reactivos más del mismo tipo general deben agregarse a la prueba para incrementar su confiabilidad a .90?

4. ¿Cuál es la diferencia entre el error estándar de medición y el error estándar de estimación? ¿Cómo se relacionan esos dos estadísticos con los coeficientes de confiabilidad y validez de una prueba?

5. ¿Cuál es el error estándar cometido al estimar los promedios académicos a partir de las calificaciones de una prueba de aptitud si la desviación estándar del criterio es .50 y la correlación entre la prueba y el criterio es .60? Si el promedio académico pronosticado de un estudiante es 2.5, ¿cuál es la probabilidad de que su promedio académico obtenido caiga entre 2.1 y 2.9? ¿Entre 1.72 y 3.28?

6. Construya una tabla empírica de expectativas para las calificaciones apareadas *X*, *Y* en la tabla A.2 del apéndice A (página 438). Deje que *X* sea la variable predictora (hilera) y *Y* la variable de criterio (columna). Use un ancho de intervalo de 7 para ambas variables al establecer los intervalos de calificación para *X* y *Y*.

7. Describa tres tipos de confiabilidad y tres tipos de validez. ¿Para qué tipos de pruebas y situaciones es más apropiado cada tipo de validez y confiabilidad?

CAPÍTULO SEIS

PRUEBAS DE APROVECHAMIENTO ESTANDARIZADAS

Las pruebas de aprovechamiento, definido como el nivel de conocimiento, habilidad o logro en un área de desempeño, son los instrumentos psicométricos más populares. Si consideramos todas las pruebas aplicadas en el salón de clases que elaboran los profesores y todas las pruebas estandarizadas vendidas a las escuelas y a otras organizaciones, el número de pruebas de aprovechamiento aplicadas sobrepasa con facilidad a todos los otros tipos de pruebas psicológicas y educativas. En Estados Unidos, la mayoría de los 50 estados ha establecido como obligatorio que los estudiantes presenten pruebas de aprovechamiento en algunos grados. La mayoría de las pruebas estandarizadas de aprovechamiento aplicadas en las escuelas estadounidenses corresponde a las áreas de lectura y lenguaje, aunque cada año se invierten también millones de dólares en pruebas de matemáticas, ciencia, ciencias sociales y otras materias.

FUNDAMENTOS DE LAS PRUEBAS DE APROVECHAMIENTO

Cualquier prueba de habilidad (inteligencia general, habilidades especiales, aprovechamiento) en realidad mide lo que la gente ha logrado. Los reactivos de las pruebas de inteligencia y habilidades especiales, como los de las pruebas de aprovechamiento, requieren que los examinados demuestren algún logro. Las calificaciones en las pruebas de aprovechamiento se utilizan para muchos de los mismos propósitos que las calificaciones en otras pruebas de habilidades generales o específicas. Esos propósitos incluyen evaluación global y diagnóstica de las habilidades del individuo, así como evaluación de la efectividad de los programas educativos y sociales.

Las pruebas de aprovechamiento educativo a menudo son mejores predictores de las notas escolares que las pruebas de inteligencia y de habilidades especiales, pero no pueden reemplazarlas por completo. Los logros medidos por las pruebas de inteligencia general son más amplios y son producto de experiencias de aprendizaje menos formales y, por lo regular, menos recientes que los logros medidos por las pruebas estandarizadas de aprovechamiento. La mayoría de las pruebas de aprovechamiento evalúa el conocimiento de algo que ha sido enseñado de manera explícita, por lo que las calificaciones en esas pruebas tienden a estar más influidas por la asesoría que las calificaciones en las pruebas de inteligencia y de habilidades especiales.

También puede hacerse una distinción entre las pruebas de aprovechamiento y otras medidas de habilidades cognoscitivas en términos de sus diferentes énfasis. Las pruebas de aprovechamiento se concentran más en el presente, es decir, en lo que la persona sabe y puede hacer ahora. Por otro lado, las pruebas de inteligencia y de habilidades especiales se concentran en el futuro: miden la aptitud para el aprendizaje, es decir, lo que una persona deberá ser capaz de hacer con educación y entrenamiento ulteriores.

Una serie de pruebas populares de aprovechamiento están vinculadas con pruebas de aptitud publicadas por la misma compañía y han sido estandarizadas en la misma población de estudiantes. El uso combinado de esas medidas de aprovechamiento y aptitud puede facilitar la interpretación de los resultados de la prueba de aprovechamiento, más allá de la información proporcionada por las normas de la prueba sola. Pueden hacerse conclusiones de si los estudiantes están desempeñándose al nivel de su potencial y en qué áreas de contenido es más probable que se beneficien de la instrucción y estudio adicionales.

Panorama histórico

Exámenes escritos en forma de composición y poesía, recopiados y juzgados por dos calificadores, se usaron por primera vez en China alrededor del año 1370 d. de C. Luego de la introducción del proceso de elaboración del papel en Europa, una habilidad que los europeos aprendieron de los árabes en el siglo XII y que éstos a su vez habían aprendido de los chinos en el siglo VIII, los exámenes escritos empezaron a reemplazar a los orales en algunas universidades europeas. Se sabe que el primer uso educativo de las pruebas escritas en una universidad europea se dio en Cambridge, Inglaterra, en 1702, y la Universidad de Londres fue acreditada como un centro de exámenes para pruebas escritas en 1836 (Green, 1991). Sin embargo, no fue sino hasta 1845 que los exámenes escritos se aplicaron a gran escala en Estados Unidos (Greene, Jorgensen y Gerberich, 1954).

A principios del siglo XIX, el número de estudiantes en las escuelas de las ciudades estadounidenses había crecido demasiado como para que la aplicación frecuente de exámenes orales resultara un recurso práctico. La examinación oral continuó siendo el principal método para evaluar el aprovechamiento de los alumnos en Estados Unidos hasta la última mitad del siglo XIX. En 1845, un educador de Boston, Horace Mann, argumentó de manera convincente que los exámenes escritos, aplicados y calificados en condiciones uniformes, eran una mejor medida del aprovechamiento que los exámenes orales. La influencia de Mann llevó a que las escuelas de Boston comenzaran a administrar cada año exámenes escritos a sus alumnos. Se esperaba que esta práctica ayudara a determinar "la condición, mejoría o deterioro de nuestras escuelas" (Fish, 1941, p. 23). A pesar de los esfuerzos de Mann y de otros educadores, durante muchos años los exámenes orales continuaron siendo el método principal para evaluar el aprovechamiento escolar y sólo gradualmente fueron reemplazados por las pruebas escritas. La calificación de las pruebas orales y escritas continuó siendo bastante subjetiva.

La primera prueba objetiva de aprovechamiento, una que podía calificarse de manera confiable, fue una escala de escritura elaborada por el inglés George Fisher en 1864. Un año después, en un esfuerzo por elevar los estándares educativos, el estado de Nueva York inició los Exámenes Regentes. Otro paso importante en la medición educativa fue dado por J. M. Rice en 1897 en su estudio clásico de las habilidades de ortografía de los escolares. Los resultados obtenidos al aplicar una prueba de ortografía de 50 palabras a 33,000 niños llevaron a Rice a concluir que se aprendía lo mismo en 15 que en 40 minutos de instrucción diaria en ortografía. En estudios posteriores, Rice elaboró pruebas objetivas para evaluar las habilidades de lenguaje y los logros aritméticos de los niños. Las pruebas de Rice por lo general se consideran como precursoras de las pruebas estandarizadas de aprovechamiento, una base sobre la que luego construyeron otros pioneros de la medición educativa.

Varias pruebas estandarizadas de aprovechamiento fueron publicadas en los primeros años del siglo XX bajo la dirección de E. L. Thorndike, a quien Ross y Stanley (1954) consideraban padre del movimiento de examinación educativa. Esas pruebas incluían la Prueba de Aritmética para Operaciones Fundamentales y la Prueba de Razonamiento Aritmético de C. L. Stone (1908), la Serie de Pruebas de Aritmética de S. A. Courtis (1909) y la Escala de Caligrafía para

Niños de Thorndike (1909). Las demostraciones de la falta de confiabilidad de las calificaciones asignadas por los maestros, incluso en las materias más exactas como matemáticas (Starch y Elliot, 1913), aumentaron el interés en las pruebas objetivas estandarizadas. Para el final de la década de 1920 se disponía de numerosas pruebas estandarizadas de aprovechamiento, incluyendo baterías de medidas como la Prueba de Aprovechamiento de Stanford (1923) para alumnos de primaria y el Examen de Contenido de Educación Superior de Iowa (1924). En 1926 la Prueba de Aptitudes Académicas de opción múltiple reemplazó a las pruebas de ensayo que previamente habían sido aplicadas por el Consejo de Examen de Ingreso a la Universidad (Donlon, 1984). El nuevo formato de opción múltiple, junto con la invención de máquinas de calificación automatizada, dio lugar a un rápido incremento en el uso de pruebas estandarizadas para la evaluación del aprovechamiento de los alumnos.

Más que haber sido motivado únicamente por intereses educativos y científicos, el crecimiento en la producción de exámenes de aprovechamiento en Estados Unidos puede atribuirse en parte al hecho de que ambos lados de un debate público sobre las escuelas públicas encontraron que la defensa y los resultados de la examinación eran políticamente útiles (Levine, 1976). Incluso hoy, la administración de pruebas estandarizadas en las escuelas sigue teniendo ramificaciones políticas significativas. El debate sobre las pruebas nacionales en las materias de educación básica (lectura, matemáticas, etc.) es ilustrativo de la política estadounidense contemporánea sobre la examinación.

Pruebas de ensayo y pruebas objetivas

A pesar de cientos de estudios de investigación, la cuestión de los méritos relativos de las pruebas de ensayo y las pruebas objetivas nunca se ha resuelto por completo. De hecho, a menudo se afirma que los maestros actuales se han excedido en el uso de las pruebas objetivas hasta llegar al detrimento de las habilidades de composición de los estudiantes. No obstante, es claro que las pruebas objetivas diseñadas con cuidado pueden medir no sólo la memorización de acontecimientos, sino también muchos de los objetivos más complejos de la instrucción que en otro tiempo se pensaba sólo podían ser evaluados mediante exámenes de ensayo. En las décadas pasadas se ha observado una tendencia notable hacia las pruebas que evalúan la obtención de objetivos instruccionales de orden superior, como la aplicación, el análisis y la evaluación. Otra tendencia ha sido la de alejarse de las pruebas estandarizadas de aprovechamiento que intentan medir el logro individual en objetivos educativos amplios y aproximarse a las pruebas diseñadas de manera específica para textos y programas de enseñanza particulares. Por último, en respuesta a la crítica de que las pruebas objetivas alientan una redacción deficiente y una autoexpresión inadecuada, ahora se concede mayor énfasis a las pruebas estandarizadas de ensayo de la expresión escrita. En un intento por ampliar la evaluación del aprovechamiento del estudiante, también se utilizan pruebas de respuesta construida en matemáticas y ciencia, protocolos de experimentos de laboratorio y portafolios del trabajo (Aiken, 1998, capítulo 5; Linn,, 1992).

Propósitos y funciones de las pruebas de aprovechamiento

La función básica de las pruebas de aprovechamiento es determinar cuánto saben las personas acerca de ciertos temas o qué tan bien pueden desempeñar ciertas habilidades. Éste es el primer propósito mencionado en la tabla 6.1. Los resultados de las pruebas de aprovechamiento informan a los estudiantes, así como a los maestros y padres, acerca de sus logros y deficiencias escolares. Otras funciones de las pruebas de aprovechamiento incluyen proporcionar información para la ubicación avanzada, la acreditación de cursos y la certificación. Esas pruebas también

TABLA 6.1 Los muchos propósitos de las pruebas de aprovechamiento

1. Evaluación de la competencia lograda
2. Diagnóstico de las fortalezas y debilidades
3. Asignación de calificaciones
4. Certificación y promoción
5. Ubicación avanzada y crédito por examinación
6. Evaluación del currículo y el programa
7. Responsabilidad
8. Información para la política educativa

Fuente: Linn, R. L. (1992). Achievement testing. En M. C. Alkin (editor), *Encyclopedia of educational research* (6ª edición, págs. 1-12. Nueva York: Macmillan)

pueden estimular el aprendizaje de los estudiantes, proporcionar a los maestros y al personal administrativo información para planificar o modificar el currículo de un estudiante o grupo de estudiantes, y servir como medio de evaluación del programa instruccional y el equipo. Las pruebas sólo miden una muestra de los logros educativos, pero se supone que esa muestra es representativa de una materia o grado particular.

Es evidente que las pruebas de aprovechamiento no son el único método para determinar la efectividad de la instrucción, pero proporcionan medidas de la calidad de la educación y, por ende, pueden contribuir a su mejoramiento. Por lo menos, las calificaciones en las pruebas de aprovechamiento sirven como señales para alertar a maestros, personal administrativo y padres acerca de las necesidades instruccionales de los estudiantes a nivel individual y colectivo (Ansley, 1997).

Las pruebas de aprovechamiento no pueden evaluar todos los objetivos o metas adoptadas por los filósofos educativos. Esas pruebas no miden de manera directa variables afectivas como el deleite y la confianza en el pensamiento, el interés en la materia educativa, el placer al usar las habilidades, el disfrute de la lectura, el aprender a aprender y a afrontar el cambio o el desarrollo de habilidades interpersonales y sociales. Lo que pueden medir, y con mayor precisión que los juicios de los maestros u otras evaluaciones subjetivas, es el grado en el que los estudiantes han alcanzado ciertos objetivos cognoscitivos de instrucción (Levine, 1976).

Pruebas donde hay mucho en juego y donde hay poco en juego

Los resultados de los exámenes pueden usarse con propósitos múltiples que conciernen tanto a individuos como a grupos. Por ejemplo, en los contextos educativos, las pruebas pueden supervisar el aprovechamiento del estudiante y evaluar la efectividad de los programas educativos. El grado en el que las decisiones aportadas por los resultados de una prueba impactan o acarrean consecuencias importantes para estudiantes y grupos se conoce como lo *que está en juego* en la prueba. Dichas decisiones pueden involucrar el diagnóstico de que un estudiante tiene una discapacidad de aprendizaje, el programa educativo apropiado para un estudiante con tal discapacidad, la ubicación de un estudiante en un programa para superdotados y talentosos, y la promoción o graduación de un estudiante de bachillerato. Otras decisiones importantes a las que contribuyen las pruebas son la admisión a cierta institución, la ubicación en un programa deseado, la obtención de una beca y la certificación o licencia profesional (Heubert y Hauser, 1999).

En contraste con las *pruebas donde hay mucho en juego,* las *pruebas donde hay poco en juego* consisten en la aplicación de un examen sólo con propósitos informativos o para juicios al-

tamente tentativos. Por ejemplo, los resultados pueden utilizarse sólo para supervisar el progreso académico y proporcionar retroalimentación sobre ese progreso a los estudiantes, maestros y padres, sin que ello implique tomar una decisión específica (American Educational Research Association *et al*., 1999).

Sea cual sea el propósito para el que puedan usarse y que estén involucradas decisiones donde hay mucho o poco en juego, es importante que todos los instrumentos psicométricos midan lo que están diseñados para medir, y que lo hagan de manera confiable. Sin embargo, cuando los resultados de una prueba se utilizan para tomar decisiones en las que hay mucho en juego y pueden tener efectos importantes en la vida de los estudiantes, es particularmente importante que la calidad de la prueba (validez, confiabilidad, estandarización y cosas similares) sea tan alta como sea posible. Debe tenerse extremo cuidado al aplicar y calificar la prueba, y los resultados deben interpretarse de manera correcta. También debe tenerse en cuenta el contexto en el cual se toman las decisiones a partir de las calificaciones.

Pruebas elaboradas por el maestro y pruebas estandarizadas

Las pruebas estandarizadas de aprovechamiento representan sólo una fracción de la cantidad de pruebas aplicadas en la escuela; los estudiantes pasan mucho más tiempo presentando pruebas elaboradas por el maestro que pruebas estandarizadas (Dorr-Bremme y Herman, 1986). Sea como sea, los propósitos o funciones de las pruebas de aprovechamiento descritos en los párrafos precedentes se aplican tanto a las pruebas administradas en el aula y preparadas por los maestros como a las estandarizadas elaboradas por profesionales en la medición educativa.

Las pruebas preparadas por el maestro difieren de las estandarizadas en ciertos aspectos importantes. Las primeras son más específicas para un maestro en particular, un salón de clases y una unidad de instrucción, y son más sencillas de mantener actualizadas que una prueba estandarizada. En consecuencia, es más probable que una prueba elaborada por el maestro refleje los objetivos educativos vigentes en una escuela o para un maestro en particular. Por otro lado, las pruebas estandarizadas se elaboran alrededor de un núcleo de objetivos educativos comunes a muchas escuelas diferentes. Esos objetivos representan los juicios combinados de expertos en la materia, quienes cooperan con los especialistas en la elaboración de pruebas para desarrollar estos instrumentos. Las pruebas estandarizadas de aprovechamiento también se interesan tanto o más en la comprensión y los procesos de pensamiento como en el conocimiento factual. De este modo, las pruebas preparadas por el maestro y las estandarizadas son complementarias más que métodos opuestos de evaluar el aprovechamiento. Miden cosas algo diferentes pero de igual importancia y, dependiendo de los objetivos del aula o escuela en particular, deben emplearse ambos tipos de pruebas. Cuando una prueba estandarizada particular no evalúa las metas educativas de cierto sistema escolar, deben considerarse otras pruebas estandarizadas o incluso una prueba elaborada por el maestro.

Además de elaborarse con mayor cuidado y de tener una cobertura de contenido más amplia que las pruebas preparadas por el maestro, las pruebas estandarizadas de aprovechamiento tienen normas y por lo general son más confiables. Por esas razones, las pruebas estandarizadas de aprovechamiento son particularmente útiles al comparar a alumnos de manera individual con el propósito de ubicación en la clase, así como en la evaluación de diferentes programas de estudio mediante la valoración de los logros relativos de escuelas y distritos diferentes. La función diagnóstica de una prueba, por medio de la cual se determinan las capacidades y discapacidades de una persona en cierta materia o área, puede ser cumplida por las pruebas preparadas por el maestro y por las estandarizadas. Sin embargo, las pruebas estandarizadas son algo más efectivas para este propósito. Las decisiones que atañen a la individualización de la enseñanza, a la ubicación de los estudiantes en niveles particulares de instrucción y a la educación terapéutica,

por lo general se toman sobre la base de las calificaciones obtenidas en pruebas estandarizadas más que en las preparadas por el maestro.

Responsabilidad

Las calificaciones de las pruebas se han empleado no sólo para evaluar el desempeño de los estudiantes, sino también para evaluar a los maestros y las escuelas. El hacer que los maestros rindan cuentas de su grado de éxito al enseñar a los estudiantes, o *responsabilidad*, ha sido un tema controvertido en la educación durante muchos años. ¿Deben los maestros, a quienes por lo general no se les permite seleccionar a sus estudiantes, pero que deben tratar de enseñar a todos los que se les asignan, ser recompensados sólo cuando alcanzan los objetivos instruccionales y no ser recompensados o incluso ser penalizados cuando no lo logran? Como resultado de la creciente preocupación pública por el fracaso de las escuelas para hacer un trabajo adecuado al educar a los estudiantes, se ha prestado particular atención a la responsabilidad por la efectividad de la enseñanza. En los sectores público y privado se han hecho intentos por responsabilizar a los maestros del aprendizaje de los estudiantes. De conformidad con esos esfuerzos, se especifican las competencias que los estudiantes deben alcanzar para completar un grado o curso de estudio o para graduarse del bachillerato. La evaluación de la efectividad de la instrucción se basa luego en la obtención de esas competencias, según lo indican en gran medida las calificaciones en las pruebas de aprovechamiento.

Por desgracia, muchos estudiantes y padres ven la educación formal desde una perspectiva más bien estrecha de vendedor-consumidor, en la cual las escuelas son vistas como mercados que "venden" productos educativos a los clientes estudiantes. Dicha perspectiva hace recaer la responsabilidad del aprendizaje del estudiante casi por completo en los maestros, los materiales educativos y la estructura y dinámicas de las organizaciones en las que tiene lugar el aprendizaje. Sin embargo, los maestros saben que es difícil, si no imposible, enseñar a estudiantes que no están interesados en aprender la materia y/o que no aceptan parte de la responsabilidad por su propia educación. De este modo, además de la responsabilidad del maestro, es necesario enfatizar la importancia de la *responsabilidad del estudiante* y de la *responsabilidad de los padres* para hacer efectivo el proceso de aprendizaje.

La siguiente carta de un maestro de octavo grado es informativa:

> Les pedí a los estudiantes de octavo grado en tres clases de matemáticas que levantaran la mano si habían planeado asistir a un colegio o universidad luego de su graduación de bachillerato. Con excepción de dos o tres estudiantes en cada grupo, todos los demás levantaron la mano. Aun así, aproximadamente la mitad de quienes dijeron que tenían planeado seguir con la educación superior no se habían molestado en terminar la tarea de matemáticas. Muchos habían estado demasiado ocupados viendo televisión, jugando videojuegos, hablando por teléfono, visitando amigos, haciendo compras o caminando por las calles en busca de algo qué hacer. En lugar de culpar a los maestros, administradores y exámenes de ingreso a la universidad por los fracasos personales, es tiempo de que los estudiantes y sus padres acepten la responsabilidad por sus éxitos o fracasos educativos. Los padres que asignan un gran valor al aprendizaje y enseñan autodisciplina, respeto por los demás, integridad personal y simplemente trabajar duro, tienen hijos con mayor probabilidad de adquirir la autoconfianza y las habilidades necesarias para lograr sus metas futuras (*US News,* 30 de abril de 2001).

Contrato de desempeño

La responsabilidad se asocia con el *contrato de desempeño*, es decir, con hacer que los salarios de los profesores se establezcan en proporción a su efectividad en la enseñanza. Un criterio importante de la efectividad en la enseñanza consiste en cambios del pretest al postest en el cono-

cimiento o la competencia del estudiante. Al usar las pruebas para determinar el grado en que los maestros han cumplido un contrato para enseñar el material educativo a los estudiantes, se aplican las mismas pruebas u otras equivalentes al inicio y al final de una unidad instruccional o un curso. En consecuencia, entre mayores sean los avances en el aprovechamiento de un estudiante del pretest al postest, mayor será el salario del maestro. Por desgracia, un resultado frecuente de la aplicación de exámenes antes y después es que se presta demasiada atención al contenido de las pruebas a expensas de otros objetivos instruccionales importantes.

Cuando se combinan con otras medidas del desempeño, las calificaciones de las pruebas de aprovechamiento pueden y deben contribuir a tomar las decisiones concernientes a la responsabilidad y el contrato de desempeño, pero tienen limitaciones definidas cuando se usan con este propósito. Puede parecer como si la determinación de la importancia de las diferencias o cambios en las calificaciones de la prueba no presentara problema. Supuestamente, todo lo que necesitamos hacer es restar las calificaciones del pretest a las del postest y analizar las diferencias de la manera que se considere apropiada. Sin embargo, un problema con este enfoque es que la diferencia en las puntuaciones crudas puede ser muy poco confiable. Esto es particularmente cierto cuando los coeficientes de confiabilidad de las calificaciones del pretest y del postest son bastante bajos, aunque sean más altos que la confiabilidad de la diferencia de las calificaciones. Otro problema estadístico encontrado al analizar la diferencia de las puntuaciones es la *regresión hacia la media*, que es la tendencia a que los examinados cuyas calificaciones en el pretest son muy bajas o muy altas obtengan en el postest calificaciones más cercanas a la media. El uso de la diferencia regresada de las calificaciones a menudo se recomienda como una forma de tratar con la regresión a la media, pero dicho procedimiento no siempre es aconsejable. Se han propuesto procedimientos estadísticos más complejos para analizar los cambios en las calificaciones de la prueba, pero todos tienen limitaciones de un tipo u otro.

Evaluaciones sumatoria y formativa

La práctica tradicional demanda aplicar una prueba de aprovechamiento al final de una unidad instruccional o de un curso para determinar si los estudiantes alcanzaron los objetivos educativos especificados. En este procedimiento, conocido como *evaluación sumatoria*, la calificación en una prueba se ve como un producto final, o suma, de unidades extensas de experiencia educativa. En contraste con la evaluación sumatoria, la necesidad de *evaluación formativa* se deriva de la creencia de que la instrucción y la evaluación deberían estar integradas. El propósito de la evaluación formativa es "ayudar tanto al aprendiz como al profesor a centrarse en el aprendizaje particular necesario para avanzar hacia el dominio" (Bloom, Hastings y Madaus, 1971, p. 61). Cuando la evaluación es formativa, las pruebas y otros métodos de evaluación del progreso educativo se aplican de manera continua durante el proceso de instrucción. Se desarrollan unidades instruccionales que incluyen los exámenes como parte integral y progresiva de la instrucción, en lugar de ser una simple culminación del proceso. De esta forma, el desempeño del aprendiz se supervisa a lo largo de la secuencia instruccional y puede servir para dirigir la revisión y el aprendizaje ulterior.

Medición con referencias a normas y a criterio

De manera tradicional, la medición educativa no sólo ha sido sumatoria más que formativa, sino que también se ha referido a normas más que a criterios. La calificación de una persona en un *prueba con referencia a normas* se interpreta comparándola con la distribución de calificaciones de un grupo de norma (estandarización) particular. Pero la calificación de una persona en una *prueba con referencia a criterio* se interpreta comparándola con un estándar o criterio estableci-

do de desempeño efectivo.[1] Este estándar puede ser formulado a partir del consenso de un grupo de personas relacionadas con todas las carreras de la vida que se interesan en la educación —profesores y personal administrativo, padres, expertos en medición y políticos. En términos del contenido, las pruebas con referencia a normas suelen ser más amplias y contener tareas más complejas que las pruebas con referencia a criterio. En consecuencia, las diferencias individuales en las calificaciones de una prueba con referencia a normas tienden a ser más extensas que las de una prueba con referencia a criterio.

A pesar de las diferencias en el propósito y diseño de las pruebas con referencia a normas y con referencia a criterio, una prueba particular de aprovechamiento puede funcionar de ambas maneras. Con frecuencia es posible determinar con el mismo instrumento cuánto material ha aprendido un estudiante (función referida a criterio) y cómo se compara su desempeño con el de otros estudiantes (función referida a normas) (Carver, 1974).

Se dispone de pruebas con referencia a criterio diseñadas para medir el aprovechamiento en una sola materia, digamos lectura o matemáticas, así como de baterías completas de estas pruebas. Otro producto ofrecido por ciertas compañías editoras de exámenes son las pruebas de una sola materia combinadas con estrategias instruccionales adecuadas para cada materia. Varias compañías dedicadas a la examinación también preparan pruebas con referencia a criterio elaboradas según ciertas especificaciones, o tienen disponibles bancos de reactivos con referencia a criterio en diversas materias. Esas pruebas elaboradas según especificaciones tienen la ventaja de estar adaptadas a los objetivos de un sistema escolar en particular, pero también tienen varias desventajas. Además del problema de decidir sobre una calificación aceptable para aprobar o el nivel de dominio en cada prueba, la necesidad de un gran número de subpruebas para medir muchos objetivos educativos diferentes requiere que cada subprueba sea relativamente corta; por ende, su confiabilidad es bastante baja. Además, no se ha resuelto del todo el problema de cómo determinar la confiabilidad y validez de las diversas subpruebas y de la prueba como un todo (Taylor y Lee, 1995).

Evaluación Nacional del Progreso Educativo

En Estados Unidos, ciertas pruebas de aprovechamiento se administran sobre una amplia base escolar, distrital o estatal para evaluar el progreso educativo de los estudiantes y supervisar la efectividad a largo plazo de programas educativos particulares. Los resultados de dicho sistema de examinación se presentan en los medios y a menudo se emplean para apoyar la acción legislativa y los gastos concernientes a la educación pública. Aunque se administra una serie de pruebas de aprovechamiento a nivel nacional, de manera periódica se efectúan pruebas distritales de aprovechamiento para evaluar el estatus educativo de muestras representativas de estudiantes en cada estado. Las pruebas administradas por la Evaluación Nacional del Progreso Educativo están próximas a merecer esta distinción.

Un enfoque con referencia a criterio ha conducido a la Evaluación Nacional del Progreso Educativo (NAEP), también conocida como La Boleta de Calificaciones de la Nación. La NAEP es un estudio continuo, a nivel nacional, del conocimiento y las habilidades, capacidades intelectuales y actitudes de los jóvenes estadounidenses. Su propósito declarado "es mejorar la efectividad de las escuelas de nuestra nación al poner a disposición de los responsables de la política a nivel nacional, estatal y local información objetiva acerca del desempeño de los estudiantes en

[1]Algunos autores (por ejemplo, Anastasi y Urbina, 1997) prefieren el término *prueba con referencia al dominio* a prueba con referencia a criterio. Ambos términos indican que el marco de referencia empleado al interpretar las calificaciones de una prueba es el contenido de la prueba, más que la muestra de examinados en los que se estandarizó ésta.

áreas selectas de aprendizaje" (Public Law 100-297, sección 3401). Desde 1969, la NAEP ha evaluado periódicamente las habilidades de grandes muestras de estadounidenses en cuatro grupos de edad (9, 13, 17 y de 25 a 35 años) en lectura, matemáticas, ciencia, redacción, historia de Estados Unidos, geografía y artes.

En la NAEP nacional se ha utilizado un procedimiento de muestreo aleatorio estratificado para seleccionar a cierto número de personas de cada género, nivel socioeconómico y raza de cuatro regiones geográficas y cuatro tipos de comunidades. Aunque se plantean muchas preguntas concernientes a cada tema, el hecho de que se muestrean tanto los examinados como los reactivos permite que sólo se necesite un periodo de prueba relativamente corto (50 minutos) por persona. A los adultos se les evalúa de manera individual, y a las personas más jóvenes tanto de manera individual como en grupo. Como los resultados se expresan en términos de los porcentajes de personas en cada grupo de edad que poseen ciertas habilidades y conocimiento, los nombres de esas personas no aparecen en los documentos de la prueba. Los resultados se presentan para la nación como un todo y para regiones geográficas específicas. Los resultados a largo plazo en matemáticas, ciencia y lectura se obtienen para las edades de 9, 13 y 17 años, y en redacción para los grados cuarto, octavo y undécimo.

Desde 1990, las evaluaciones de la NAEP también se han realizado de manera voluntaria a nivel estatal. Se seleccionan muestras separadas representativas de estudiantes para cada jurisdicción o estado participante, pero los resultados no son representativos del estado en general.

La NAEP fue planificada como un programa continuo para proporcionar al público estadounidense, y en especial a los legisladores y educadores, información sobre el estado y crecimiento de los logros educativos en Estados Unidos y sobre el grado en que se están alcanzando las metas educativas de esa nación. No fue diseñada, como algunos han temido, para evaluar los logros de escuelas o distritos escolares específicos o como un medio de control federal sobre los programas de las escuelas públicas. Sin embargo, los hallazgos han sido analizados por área geográfica, tamaño y tipo de comunidad, género, educación de los padres y grupo étnico. De particular interés son los análisis de los efectos del apoyo federal y de tipos específicos de programas sobre los logros educativos.[2]

TIPOS Y SELECCIÓN DE LAS PRUEBAS DE APROVECHAMIENTO ESTANDARIZADAS

Existen cuatro tipos de pruebas de aprovechamiento estandarizadas: baterías de pruebas de estudio, pruebas de estudio en materias especiales, pruebas de diagnóstico y pruebas de pronóstico. Algunas son pruebas individuales diseñadas para aplicarse a una persona a la vez, pero la gran mayoría son pruebas colectivas que pueden aplicarse a cualquier número de personas al mismo tiempo. El mercado para pruebas muy especializadas en un área temática particular es más bien limitado, por lo que las pruebas estandarizadas de aprovechamiento por lo regular cubren áreas amplias de contenido y tratan con materias de conocimiento general. Debido a que el currículo se vuelve más especializado en los niveles superiores, la administración de pruebas estandarizadas de aprovechamiento es menos común después de la secundaria.

[2]Es posible obtener informes de la NAEP y publicaciones relacionadas en ED Pubs, P.O. Box 1398, Jessup, MD 20794-1398. Teléfono: 877-4ED-PUBS. FAX: 301-470-1244. Direcciones Web: http://www.ed.gov/pubs/edpubs.html y http://nces.ed.gov/nationsreportcard.

Baterías de pruebas de estudio

La forma más integral de evaluar el aprovechamiento es aplicando una batería de pruebas de estudio, que es un conjunto de pruebas sobre una materia diseñadas para un nivel particular. El propósito principal de aplicar una batería de pruebas es determinar la posición general de un individuo en varias materias, más que medir sus fortalezas y debilidades específicas. En consecuencia, cada prueba de una batería de estudio contiene una muestra bastante limitada del contenido y las habilidades de una materia en particular. Como todas las pruebas de una batería se estandarizan en el mismo grupo de personas y las calificaciones se expresan en la misma escala numérica, el desempeño de una persona en diferentes materias puede compararse de manera directa.

Aunque las baterías de pruebas proporcionan una evaluación más amplia del aprovechamiento de los alumnos que las pruebas sencillas, tienen una serie de desventajas. A pesar de que el tiempo total de administración de una batería es más largo, las pruebas son más cortas que las pruebas de estudio sencillas por lo que su confiabilidad suele ser menor. Por supuesto, no es necesario administrar todas las pruebas de una batería a un grupo dado de estudiantes; el examinador puede decidir administrar sólo las pruebas que proporcionen información relevante relacionada con las metas específicas de la evaluación.

Pruebas de estudio de una sola materia

Las pruebas de una sola materia por lo general son más largas y más detalladas que las pruebas comparables en una batería, por lo que permiten una evaluación más pormenorizada del aprovechamiento en un área específica. Las pruebas de una sola materia arrojan regularmente una calificación global y quizás un par de subcalificaciones, y no fueron diseñadas para identificar causas específicas de alto o bajo desempeño en la materia. Debido a la mayor uniformidad existente entre las diferentes escuelas en lo que toca a la instrucción de la lectura y las matemáticas más que en otras materias, las pruebas estandarizadas en esas dos áreas tienden a ser más válidas que, por ejemplo, las pruebas en ciencia y ciencias sociales.

Pruebas de diagnóstico

Estas pruebas tienen la función diagnostica de identificar dificultades específicas en el aprendizaje de una materia. Para elaborar una prueba de diagnóstico en una habilidad básica como lectura, aritmética u ortografía, se analiza el desempeño en la materia como un todo en subhabilidades, y luego se elaboran grupos de reactivos para medir el desempeño en esas subhabilidades. A diferencia de las pruebas de estudio, que se concentran en las calificaciones totales, las pruebas de diagnóstico generan calificaciones en cada una de varias subhabilidades. Como las diferencias entre calificaciones en las diversas partes de las pruebas se interpretan al hacer diagnósticos, el número de reactivos para medir una subhabilidad particular debe ser suficiente para asegurar que las diferencias entre las calificaciones de las partes sean confiables. Por desgracia, el número de los reactivos que componen las calificaciones de las partes a menudo es pequeño y las calificaciones de las partes se correlacionan, lo que da por resultado que las diferencias de las calificaciones tengan poca confiabilidad.

La mayoría de las pruebas de diagnóstico son de lectura, pero también se dispone de estas pruebas en matemáticas, ortografía y lenguas extranjeras. Una prueba de diagnóstico contiene una mayor variedad de reactivos y, por lo general, su administración se lleva más tiempo que una prueba de estudio de la misma materia. Las pruebas de diagnóstico también pueden implicar el uso de aparatos especiales, como un taquitoscopio, para presentar el material de lectura sólo por

un periodo breve, y la cámara de movimientos oculares para seguir la dirección en que se mueven los ojos al leer.

Ciertas pruebas de estudio de administración individual, o pruebas *globales*, también se utilizan con propósitos de diagnóstico educativo. Algunos ejemplos son la Prueba de Aprovechamiento Educativo de Kaufman y la Prueba de Aprovechamiento Individual de Peabody, Revisada. Aún más globales en sus propósitos de diagnóstico son las Pruebas de Aprovechamiento de Woodcock-Johnson III, una batería de pruebas de habilidades múltiples de administración individual diseñada para medir la habilidad intelectual general, habilidades cognoscitivas específicas, lenguaje oral y aprovechamiento académico de individuos de entre 2 y 90 años de edad.

La administración de una batería de pruebas de estudio es un primer paso razonable en un programa de examinación porque proporciona una imagen global de la posición de una persona en varias materias. Si se necesita una segunda evaluación del aprovechamiento de una persona en un área particular, puede administrarse una sola prueba de la materia específica. Por último, si se requiere hacer un análisis detallado de la discapacidad de una persona en lectura o matemáticas y determinar las causas de la discapacidad, debe administrarse una prueba de diagnóstico.

Pruebas de pronóstico

Las pruebas de pronóstico, al igual que las pruebas de aptitud, contienen una mayor variedad de reactivos que las pruebas de estudio del aprovechamiento en la misma materia, ya que están diseñadas para predecir el aprovechamiento en materias escolares específicas. Por ejemplo, el propósito de una prueba de preparación para la lectura aplicada a un alumno de jardín de niños o de primer grado es predecir si el niño está preparado para beneficiarse de la enseñanza de la lectura. A un nivel superior, se dispone de pruebas de pronóstico en matemáticas (álgebra, geometría) y en lenguas extranjeras con el fin de predecir la facilidad para el aprendizaje de esas materias.

Selección de una prueba estandarizada

La selección de una prueba estandarizada de aprovechamiento básicamente es cuestión de encontrar un instrumento con un contenido que se ajuste a los objetivos instruccionales de una organización, clase, escuela o sistema escolar particular. Esto significa que el nivel de conocimiento o habilidad de los examinados y el contenido y objetivos del currículo deben determinarse antes de decidir qué prueba(s) administrar. Además, deberán considerarse las razones para administrar la prueba y la forma en que van a usarse las calificaciones; no tiene sentido administrar una prueba simplemente porque "parece buena" y luego dejar que los resultados no utilizados se empolven en una gaveta o en un armario.

Propósitos y consideraciones prácticas. El manual que acompaña a una prueba por lo regular proporciona detalles sobre sus usos potenciales (evaluación, ubicación, diagnóstico de las discapacidades de aprendizaje, preparación para aprender, evaluación del currículo) y cita evidencia de apoyo. En consecuencia, antes de seleccionar una prueba deben aclararse las formas específicas en que van a usarse las calificaciones y consultarse los manuales de la prueba para determinar qué instrumentos son apropiados para esos propósitos. Además de leer el manual, los posibles usuarios deben examinar una copia de la prueba e incluso resolverla para determinar si es adecuada para sus propósitos. Algunas empresas también publican muestras de las pruebas que editan, las cuales constan de un folleto de la prueba, una hoja de respuestas, un manual, una clave de calificación y otros materiales asociados. También pueden solicitarse catálogos de pruebas. Esos materiales son útiles para decidir qué pruebas administrar. La mayoría de las com-

pañías de pruebas también tienen sitios Web en los que describen sus propósitos, productos y servicios (vea el apéndice C).

Otra cosa que debe considerarse al seleccionar una prueba es el grado de cooperación que puede esperarse de la escuela u otra organización al administrarla e interpretar los resultados. También son de importancia cuestiones prácticas como costo y tiempo de aplicación, calificación y análisis de los resultados. Los servicios de calificación por medio de una máquina proporcionados por firmas comerciales de pruebas facilitan en gran medida los procesos de calificación y análisis y, por lo común, son de un costo bastante razonable.

Confiabilidad, validez y normas. Las características estadísticas de las pruebas de aprovechamiento suelen pasarse por alto al momento de seleccionar una prueba de este tipo, pero es crucial atender este aspecto. La confiabilidad de la mayoría de las pruebas de aprovechamiento se ubica entre .80 y .90, pero el significado de esos altos coeficientes depende de los procedimientos con que se obtuvieron. Un coeficiente de formas paralelas es preferible a un coeficiente de test-retest o a uno de consistencia interna porque es más probable que los dos últimos estén inflados por el error de medición. Para decidir si una prueba de aprovechamiento es válida, debe obtenerse evidencia de su validez de contenido comparando éste con los objetivos del programa instruccional de interés. Un manual de la prueba preparado adecuadamente describe el sistema para clasificar el contenido y los objetivos conductuales utilizados al elaborar la prueba, y los usuarios potenciales deben decidir si esos objetivos corresponden a los suyos. Cuando se administra una prueba con el propósito de predecir el aprovechamiento posterior, como sucede con una prueba de preparación para la lectura u otra prueba de pronóstico, también es importante obtener evidencia de su validez predictiva.

Además de la confiabilidad y la validez, antes de seleccionar una prueba también debe examinarse si las normas son adecuadas y apropiadas. La mayoría de las pruebas de aprovechamiento bien elaboradas se estandarizaron en muestras (estadounidenses) nacionales representativas, en ocasiones estratificadas de acuerdo con edad, sexo, región geográfica, posición socioeconómica y otras variables relevantes. Los compradores de la prueba que planean presentar las calificaciones en términos de esas normas deben asegurarse de que las características del grupo de norma son similares a las de los estudiantes que van a examinarse. Para propósitos de ubicación y otras comparaciones dentro de una escuela o sistema escolar determinado, las normas locales pueden ser incluso más significativas que las nacionales.

Los usuarios de las pruebas estandarizadas de aprovechamiento también deben estar al tanto de que, al trazar el progreso académico de un estudiante por medio de calificaciones normadas en una prueba estandarizada de aprovechamiento aplicada a niveles sucesivos, se asume que los grupos de diferentes niveles en los que se estandarizó la prueba son equivalentes. Por ejemplo, los cambios demográficos en las comunidades de las que se extrajeron estudiantes de ciertas escuelas pueden producir diferencias significativas en la composición de grupos de estudiantes de diferentes niveles. Esto puede suceder debido a la llegada migratoria reciente de personas que difieren en el nivel socioeconómico, nacionalidad o grupo étnico. Si hay razones para creer que existen diferencias significativas entre los grupos de norma en variables distintas a las relacionadas con el crecimiento, entonces las calificaciones normadas por grado, de rango percentilar o estándar obtenidas por un estudiante en una prueba no pueden compararse con precisión entre los niveles.

Al adquirir una prueba es importante no dejarse engañar por su nombre. Los usuarios de pruebas experimentados están bien conscientes de que es un error suponer que instrumentos con el mismo nombre miden la misma cosa e instrumentos que tienen nombres diferentes miden cosas distintas. Antes de decidir qué pruebas de aprovechamiento adquirir, tanto los usuarios no-

vatos como los experimentados pueden beneficiarse de consultar *The Mental Measurements Yearbook, Test Critiques* y las revisiones de pruebas en revistas profesionales y otras fuentes.

BATERÍAS DE PRUEBAS DE APROVECHAMIENTO

Las baterías de pruebas de aprovechamiento representan esfuerzos por medir las amplias capacidades y habilidades cognoscitivas cultivadas por las experiencias educativas en áreas centrales. Estas baterías de pruebas de niveles múltiples evalúan destrezas básicas en lectura, matemáticas, lenguaje y, a los niveles apropiados, habilidades de estudio, ciencias sociales y ciencia.

Es posible encontrar descripciones de baterías de pruebas de aprovechamiento que están comercialmente disponibles en las diversas ediciones de *The Mental Measurements Yearbook, Tests in Print, Tests* y *Test Critiques*, así como en los catálogos de los editores de pruebas. Tales baterías fueron diseñadas para evaluar el aprovechamiento educativo formal de estudiantes desde el jardín de niños hasta bachillerato, con énfasis en los años de primaria y secundaria.

Los programas de exámenes de muchas escuelas se basan en las baterías de pruebas de aprovechamiento aplicadas en otoño y primavera a sus alumnos con el propósito de medir el logro y el progreso educativo general. Los resultados de estas pruebas son de interés para los maestros, padres, personal administrativo, miembros de los consejos escolares, líderes políticos y, por supuesto, para los estudiantes. Una limitación del uso de baterías es que algunas de las pruebas pueden no corresponder a los objetivos particulares de la escuela o sistema escolar. Además, no todas las pruebas en una batería determinada tienen igual confiabilidad o la misma validez de contenido.

Normas de una batería de pruebas

Debido a que las diversas subpruebas de un nivel particular en una batería de pruebas de aprovechamiento se estandarizaron en el mismo grupo de personas, el conjunto unificado de normas resultantes permite la evaluación directa del aprovechamiento relativo de una persona en varias áreas temáticas. Además, si puede asumirse que diferentes niveles de una batería de pruebas se estandarizaron en grupos comparables de estudiantes, entonces el progreso cognoscitivo del alumnado puede trazarse comparando sus calificaciones en las pruebas que componen la batería a lo largo de un periodo de varios años. Sin embargo, esto no debe hacerse cuando existe alguna duda acerca de la equivalencia o posibilidad de comparación de las diferentes muestras de nivel de los estudiantes en los que se estandarizó la batería. Además, las normas contra las que se comparan las calificaciones de los estudiantes deben haberse obtenido de la aplicación de la(s) prueba(s) al grupo de estandarización en la misma época del año (otoño o primavera) en que se examine a los alumnos cuyas calificaciones están siendo evaluadas.

Contenido de las baterías de pruebas de aprovechamiento

Nivel de escuela primaria. Debido a la mayor uniformidad del contenido instruccional en la primaria, las baterías de pruebas de aprovechamiento se administran con mayor frecuencia en este nivel para evaluar el desarrollo educativo. Una batería típica para la escuela primaria consta de subpruebas sobre vocabulario de lectura, lectura de comprensión, uso del lenguaje, ortografía, aritmética básica y comprensión de la aritmética. También puede incluir subpruebas para medir habilidades de estudio, ciencias sociales y ciencia, pero al nivel de primaria se enfatiza la medición del aprovechamiento en habilidades cuantitativas y verbales básicas. Las baterías populares de pruebas de aprovechamiento para este nivel incluyen la Serie de Pruebas de Aprovechamiento de Stanford, las Pruebas de Aprovechamiento de California, la Prueba Comprensiva

de Habilidades Básicas y las Pruebas de Aprovechamiento Metropolitanas. Esas baterías también contienen pruebas para niveles de jardín de niños y secundaria.

Nivel de escuela secundaria. Debido a la variabilidad en los programas académicos de diferentes estudiantes de nivel medio, las baterías de pruebas de aprovechamiento son menos útiles a este nivel. Las baterías de pruebas al nivel de escuela secundaria siguen enfatizando las habilidades básicas en lectura, lenguaje y aritmética, pero también se incluyen pruebas de ciencias sociales, ciencia y habilidades de estudio. Tanto a nivel de primaria como de secundaria, las pruebas de aprovechamiento enfatizan el desarrollo educativo general y no están vinculadas a cursos específicos en escuelas particulares. Al nivel de la educación media también son de interés baterías como las Pruebas Universitarias Estadounidenses (ACT), las cuales se administran anualmente con propósitos de admisión a la universidad. La ACT es en realidad una batería de pruebas de aprovechamiento, pero es similar a una prueba de aptitud en el hecho de que su amplio rango de contenido se relaciona menos con experiencias escolares específicas que la mayoría de las pruebas de aprovechamiento.

Pruebas de educación básica

Varias baterías de pruebas de aprovechamiento se han diseñado de manera específica para medir la competencia en las habilidades básicas de los adultos con educación inferior al nivel medio. Un ejemplo son las Pruebas de Educación Básica para Adultos (TABE) (de CTB/McGraw-Hill), las cuales constituyen una prueba de niveles múltiples estandarizada en adultos que destaca las habilidades en lectura, matemáticas y lenguaje. Otra prueba para determinar el nivel de desarrollo en lectura y aritmética de empleados o solicitantes en una amplia variedad de ocupaciones y ambientes de rehabilitación es el Índice de Lectura-Aritmética (RAI) (de NCS London House). En la figura 6.1 se presentan reactivos de muestra de esta prueba, la cual, si bien no se cronometra, se lleva alrededor de 25 minutos por cada una de sus dos partes.

A pesar de la disponibilidad de pruebas de habilidades básicas para adultos, sólo una minoría de los negocios y las industrias evalúan en realidad la alfabetización de sus empleados. En consecuencia, muchos trabajadores son funcionalmente iletrados y deben "engañar" al realizar un trabajo que requiere habilidades de lectura. Es de suponer que los ejecutivos de dichas compañías se dan cuenta de que algunos de sus empleados no pueden leer, escribir, realizar cálculos o comprender bien el idioma, pero parecen estar limitados en lo que pueden hacer acerca de esta situación. Esto es desafortunado porque los empleados analfabetas tienen mayor probabilidad de sufrir accidentes y se ven impedidos en su capacidad para avanzar en una organización.

Pruebas GED

Las Pruebas de Desarrollo Educativo General (GED) (de GED Testing Service) también son apropiadas para adultos con educación formal limitada, y son presentadas cada año por más de 800,000 adultos. Las pruebas GED fueron diseñadas para medir los logros educativos de personas con educación media o equivalente. La batería completa, que se lleva alrededor de siete horas y media, consta principalmente de reactivos de opción múltiple en cinco áreas: habilidades de redacción, ciencias sociales, ciencia, literatura y arte, y matemáticas. La prueba de habilidades de redacción también incluye un ensayo que documenta la habilidad del examinado para escribir y comunicarse de manera efectiva. En lugar de enfatizar hechos y detalles específicos, los reactivos de la GED tratan sobre conceptos amplios y generalizaciones basadas en competencias y conocimiento enseñados en los programas académicos de la secundaria. Muchas organizaciones académicas y de negocios, así como las fuerzas armadas de Estados Unidos, aceptan califi-

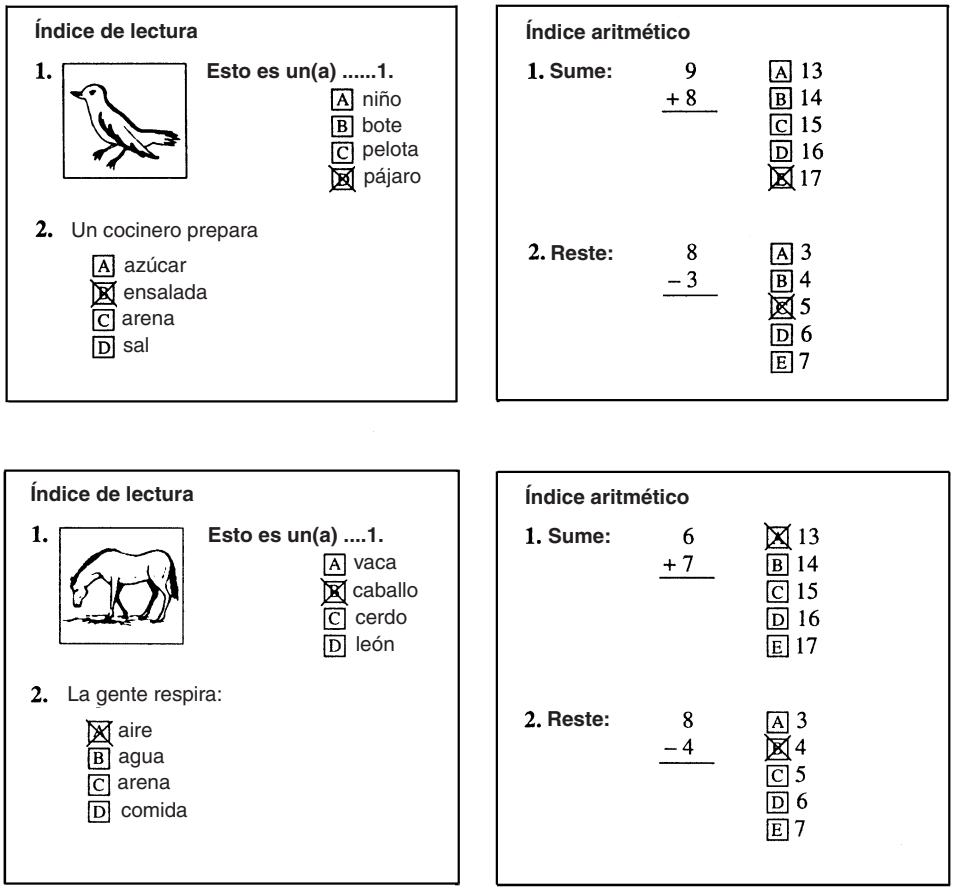


FIGURA 6.1 Muestra de reactivos del Índice de Lectura-Aritmética.

caciones en esas pruebas de *diploma de equivalencia general* sobre la misma base que el diploma de secundaria (vea el sitio Web www.gedtest.org).

PRUEBAS DE APROVECHAMIENTO EN ÁREAS ESPECÍFICAS

La aplicación de una batería de pruebas de aprovechamiento tiene prioridad en un programa escolar de pruebas típico. Cuando se necesita más información sobre el desempeño del estudiante en una materia particular, el procedimiento usual es administrar una prueba específica en esa materia luego de la batería. Esas pruebas específicas de aprovechamiento tienen ciertas ventajas sobre pruebas comparables en una batería. Por ejemplo, el que una prueba específica contenga más reactivos y una temática más amplia que la prueba de una batería de aprovechamiento, le da mayor probabilidad de representar de manera más adecuada los objetivos instruccionales de una amplia gama de aulas y de escuelas. Además, debido a su extensión, probablemente sea más confiable que una prueba comparable en una batería de aprovechamiento.

Una línea de una antigua canción inglesa, "Reading and writing and 'rithmetic, taught to the tune of a hickory stick", es un testimonio de la relevancia que han tenido esas materias a lo largo del tiempo en el programa de estudios de la escuela primaria. Se dispone de cientos de pruebas para materias específicas en lectura, matemáticas, lenguaje, ciencia, ciencias sociales, profesiones, negocios y oficios. Otras áreas en las que se han publicado pruebas estandarizadas de aprovechamiento son salud, economía doméstica, artes industriales, uso de las bibliotecas, literatura, la Biblia, música, oratoria, ortografía y educación vial. Además de las pruebas tradicionales con referencia a normas del tipo de estudio, diagnóstico y pronóstico, hay muchas pruebas con referencia a criterio en materias específicas. Más aún, el énfasis que en las décadas recientes se dio en las secundarias a las pruebas de competencia en habilidades básicas llevó a la publicación de una serie de pruebas de competencia para evaluar el conocimiento y las habilidades de estudiantes de secundaria y preparatoria en lectura, redacción y matemáticas. Esas habilidades de supervivencia se consideran esenciales para enfrentar las demandas de la vida diaria.

Pruebas de lectura

Muchas de las dificultades experimentadas por los niños en el aprendizaje de las materias escolares se relacionan con problemas en la lectura, una razón común para canalizar a un niño a evaluación psicoeducativa. Las dificultades en la lectura son acumulativas y afectan el desempeño en casi todo el trabajo escolar, por lo que es importante evaluar el nivel de lectura y diagnosticar deficiencias en esta materia de manera oportuna y regular. Debido a sus muchos usos, se administran más pruebas de lectura que cualquier otro tipo de prueba de aprovechamiento. Se dispone de varios tipos de pruebas de lectura, siendo las tres categorías principales las pruebas de estudio, pruebas de diagnóstico y pruebas de preparación para la lectura. Otras formas de clasificar las pruebas de lectura son con referencia a norma y a criterio (o ambas) y lectura en silencio y lectura oral.

Pruebas de estudio de lectura. La razón principal para aplicar una prueba de estudio de lectura es determinar la habilidad general de una persona para la lectura. Las pruebas de este tipo contienen secciones de reactivos de vocabulario y secciones de párrafos o pasajes acerca de los cuales se plantean preguntas. Se obtiene una medida del conocimiento de las palabras a partir de los reactivos de vocabulario, mientras que la velocidad y el nivel de comprensión se miden a partir de los párrafos. Algunos ejemplos de las mejores pruebas de este tipo son las Pruebas de Lectura de Gates—MacCinitie (GMRT), cuarta edición. Diseñadas para los grados K—12 y Lectura de Adultos, las dos formas (S y T) de la GMRT contienen cinco niveles: Prelectura (PL), Lectura de Principiantes (LP), 1 y 3, 3—12 y Lectura de Adultos (LA). Las habilidades de lectura de principiantes y de nivel primaria se evalúan en los niveles inferiores, y el progreso continuo en la competencia para la lectura se mide en los niveles superiores.

La mayoría de las pruebas de estudio de lectura emplean un formato de respuesta de opción múltiple, pero en la Prueba de Lectura Stanford 9 de Final Abierto se utiliza un formato abierto-cerrado o de respuesta elaborada. Otros dos ejemplos de pruebas de estudio de lectura son la Prueba de Lectura Oral de Gray, revisada, y la Prueba de Comprensión de Lectura (de pro.ed). Algunas pruebas de estudio de lectura, como el CD-ROM de la Prueba de Lectura de Nelson-Denny, pueden administrarse por medio de una computadora.

Pruebas de diagnóstico de lectura. Las pruebas de diagnóstico de lectura, que son por mucho el tipo más común de pruebas de diagnóstico, pretenden evaluar muchos factores diferentes que afectan la lectura y, por ende, descubrir la fuente de las discapacidades de los estudiantes en la materia. Entre esos factores se incluyen la coordinación ojo-mano, la percepción visual y audi-

tiva, la comprensión de conceptos e incluso la motivación. Una prueba de diagnóstico de lectura puede contener subpruebas en discriminación visual y auditiva, vocabulario de vista y vocabulario en contexto, fonemas/grafemas, vocales y consonantes, lectura en silencio y oral, lectura de comprensión y tasa de comprensión. Como las correlaciones entre esas subpruebas a menudo son sustanciales, las diversas habilidades medidas por las pruebas de diagnóstico de lectura no son necesariamente independientes. Además, la confiabilidad de las subpruebas y de la prueba como un todo frecuentemente no es tan alta como sería deseable. Algunas pruebas representativas de esta categoría son las Pruebas de California para el Diagnóstico de la Lectura (de CTB/McGraw-Hill), las Pruebas de Stanford para el Diagnóstico de la Lectura, cuarta edición (de Harcourt Brace) y la Batería de Diagnóstico de la Lectura de Woodcock (de Riverside Publishing).

Pruebas de preparación para la lectura. Como medida del grado en que los niños poseen las habilidades y el conocimiento necesarios para aprender a leer, una prueba de preparación para la lectura con frecuencia permite formular una mejor predicción del aprovechamiento en primer grado que una prueba de inteligencia general, y requiere menos tiempo de aplicación. Las pruebas de preparación para la lectura contienen muchos de los mismos tipos de reactivos que las pruebas de diagnóstico de lectura, y ciertas pruebas de lectura contienen componentes de diagnóstico y de pronóstico.

Pruebas de matemáticas

De manera similar a las pruebas de aprovechamiento en lectura, las de aprovechamiento en matemáticas pueden clasificarse como de estudio, diagnóstico y pronóstico.

Pruebas de estudio de matemáticas. Diversos enfoques hacia la instrucción están representados por las pruebas actuales de matemáticas, incluyendo el énfasis más tradicional en los programas de matemáticas así como puntualizaciones más modernas en lo relativo a resolución de problemas, desarrollo de conceptos y razonamiento. Ciertas pruebas están diseñadas para abarcar los énfasis moderno y tradicional en los programas de matemáticas, y se dispone de instrumentos que reflejan enfoques instruccionales más especializados desde el nivel de primaria hasta el de universidad. En general, las pruebas de matemáticas con referencia a normas del tipo de estudio requieren que los estudiantes demuestren cierta comprensión de conceptos y operaciones cuantitativas y la habilidad para aplicar esta comprensión a la resolución de problemas. Las pruebas de competencia en cursos generales y específicos de matemáticas (álgebra, cálculo, trigonometría) a nivel de secundaria se encuentran disponibles en el Programa de Exámenes de Nivel Universitario (CLEP).

Pruebas de diagnóstico en matemáticas. Aunque se aplican menos que las pruebas de diagnóstico para la lectura, las pruebas de diagnóstico en matemáticas también representan intentos por descomponer una materia compleja que involucra una variedad de habilidades en los elementos que la constituyen. Los reactivos en las pruebas de diagnóstico de aritmética y matemáticas se basan en un análisis de habilidades y errores en la materia. Esas pruebas incluyen el conocimiento y las habilidades requeridos para aplicaciones que involucran numeración, fracciones, álgebra y geometría. Dos ejemplos de pruebas de diagnóstico en matemáticas son la Prueba de Stanford para el Diagnóstico en Matemáticas, cuarta edición (de Harcourt Brace) y la KeyMath, Revisada/NU: Un Inventario de Diagnóstico de Matemáticas Esenciales (de American Guidance Service). El primer instrumento es una prueba de grupo diseñada para diagnosticar las fortalezas y

debilidades específicas en conceptos y operaciones matemáticas básicas de niños de primero a doceavo grado. KeyMath es una prueba de administración individual diseñada para medir la comprensión y aplicación de los conceptos y habilidades matemáticas básicas desde el jardín de niños hasta el noveno grado.

Pruebas de pronóstico en matemáticas. Se han diseñado varias pruebas para pronosticar el desempeño en cursos específicos de matemáticas, pero en comparación con las pruebas de pronóstico de la lectura (pruebas de preparación para la lectura), no son de uso común. Dos ejemplos son la Prueba de Pronóstico en Álgebra de Orleans-Hanna, tercera edición (de Harcourt Brace) y la Prueba de Aptitud para el Álgebra de Iowa, cuarta edición (de Riverside Publishing). Diseñada para identificar qué estudiantes tendrán éxito y cuáles enfrentarán dificultades al aprender álgebra, la Orleans-Hanna evalúa aptitud y aprovechamiento, así como el interés y la motivación para el álgebra, de estudiantes de secundaria y preparatoria. Se necesitan 40 minutos para resolver el cuestionario y los reactivos de la muestra de trabajo de la prueba. El rango percentilar y las normas del tipo estaninas se basan en tres grupos de estudiantes: los que terminaron matemáticas de séptimo grado, los que terminaron matemáticas de octavo grado y aquellos de los dos primeros grupos que terminaron un curso de un año en álgebra en el año siguiente. La Prueba de Aptitud para el Álgebra de Iowa fue diseñada para evaluar la preparación en Álgebra I de los estudiantes de séptimo y octavo grados. Sus cuatro subpruebas, cuya solución requiere un total de 50 minutos, miden las habilidades de pre-álgebra al interpretar gráficas e información matemática escrita, la traducción de problemas en palabras a un formato algebraico o de ecuaciones, la identificación de funciones y el uso de símbolos.

Pruebas de lenguaje

El lenguaje, tal como suele interpretarse el término, se refiere a cualquier forma de comunicación. Aunque las pruebas de lenguaje consisten principalmente en reactivos de tipo verbal, se han desarrollado medidas de comunicación no verbal para usar con personas que tienen problemas de audición e incluso con personas de audición normal. El lenguaje oral y el escrito se enseñan en todos los niveles y se dispone de pruebas apropiadas para todos los grados. El fracaso para entender ciertos conceptos puede actuar como barrera e impedir la comunicación entre los alumnos de preprimaria y primaria y los maestros, y en consecuencia afectar seriamente el aprendizaje de los niños. Como reconocimiento de este hecho, se diseñaron la Prueba Boehm de Conceptos Básicos, tercera edición (para grados K-2) y la Boehm-3 (para edades de tres a cinco años) para medir el dominio que tiene un niño pequeño de los conceptos básicos de espacio, cantidad y tiempo (vea la figura 6.2).

A pesar de la disponibilidad de pruebas como la Boehm, la mayoría de las pruebas de aprovechamiento en la categoría de lenguaje se diseñó para estudiantes de secundaria y universidad. Esos instrumentos, que incluyen pruebas en inglés y lenguas extranjeras, con frecuencia se aplican en el bachillerato y en las universidades con el propósito de colocar a los estudiantes en cursos de inglés o de lenguas extranjeras de acuerdo con su nivel de competencia.

Pruebas del idioma inglés. Algunas de las críticas más severas a las pruebas objetivas han venido de maestros de inglés, pero por lo general se reconoce que desempeñan un buen trabajo en la medición del conocimiento de gramática y vocabulario, y, en cierto grado, de las habilidades en la expresión oral y escrita. La evaluación de las habilidades en el idioma inglés forma parte de las baterías de pruebas de aprovechamiento, pero también existen otras muchas pruebas distintas para medir la competencia en inglés.

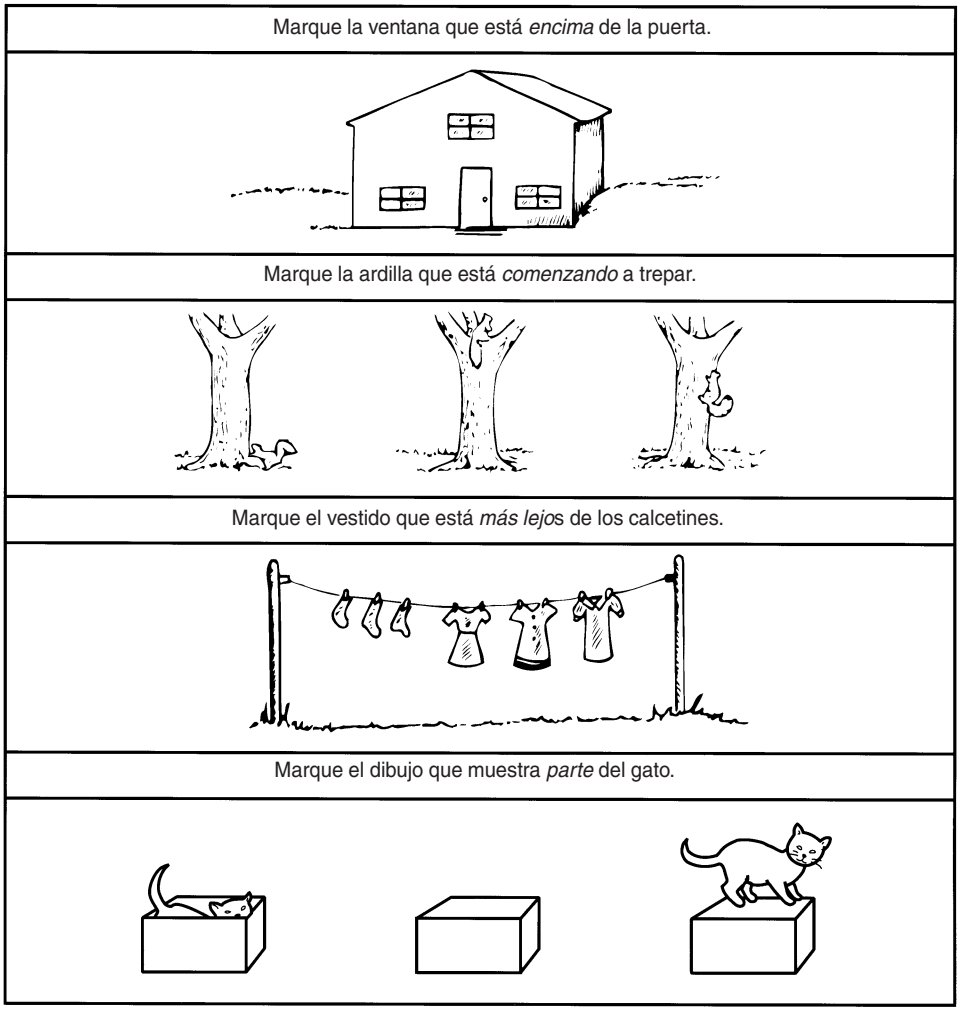


FIGURA 6.2 Muestra de reactivos de la Prueba Boehm de Conceptos Básicos, tercera edición.
El examinado marca con una × la opción seleccionada.

Como es evidente, escuchar, hablar y escribir forman parte del uso del inglés y se ha diseñado una serie de pruebas para medir esas habilidades. Ejemplo de una prueba de este tipo es la serie OWLS: Escala de Comprensión Auditiva, Escala de Expresión Oral y Escala de Expresión Escrita (de American Guidance Service). La resolución de cada una de esas pruebas, las cuales son apropiadas para niños y adultos jóvenes, se lleva menos de 25 minutos. La Escala de Comprensión Auditiva mide el lenguaje receptivo, la Escala de Expresión Oral mide el lenguaje expresivo y la Escala de Expresión Escrita proporciona una evaluación auténtica de las habilidades del lenguaje escrito. Las habilidades de hablar y escuchar en inglés o español pueden medirse

con las Escalas de Evaluación del Lenguaje Oral (LAS-O) y las Pre-LAS 2000 (de CTB/McGraw-Hill). Las escalas LAS-O se aplican de primero a duodécimo grados y las Pre-LAS a niños preescolares.

Algunos ejemplos de pruebas de escritura son la Prueba de Lenguaje Escrito-3 (TOWL-3) (de pro.ed) y el Programa de Stanford de Evaluación de la Escritura, tercera edición (de Harcourt Brace). Diseñada para estudiantes de segundo a duodécimo grados, la TOWL-3 es una medida de muestra de trabajo de respuesta libre en la cual el examinado escribe historias acerca de cierto conjunto de imágenes (vea la figura 6.3). Las historias pueden calificarse en varias variables, incluyendo tema, vocabulario, sintaxis, ortografía y estilo. La Evaluación Stanford de Escritura implica la presentación de una serie de sugerencias escritas diseñadas para provocar determinada muestra de escritura en cada uno de cuatro modos descriptivos: descriptivo, narrativo, expositor y persuasivo. Una Lista de Verificación del Escritor proporciona recordatorios para elaborar un borrador, componerlo y editarlo. La escritura se califica en ideas y desarrollo, organización, unidad y coherencia; frases y párrafos; gramática y uso, y mecánica.

Muchas otras pruebas de aprovechamiento como las Pruebas de Ubicación Avanzada CEEB y los Exámenes del Registro de Graduados también contienen un componente escrito

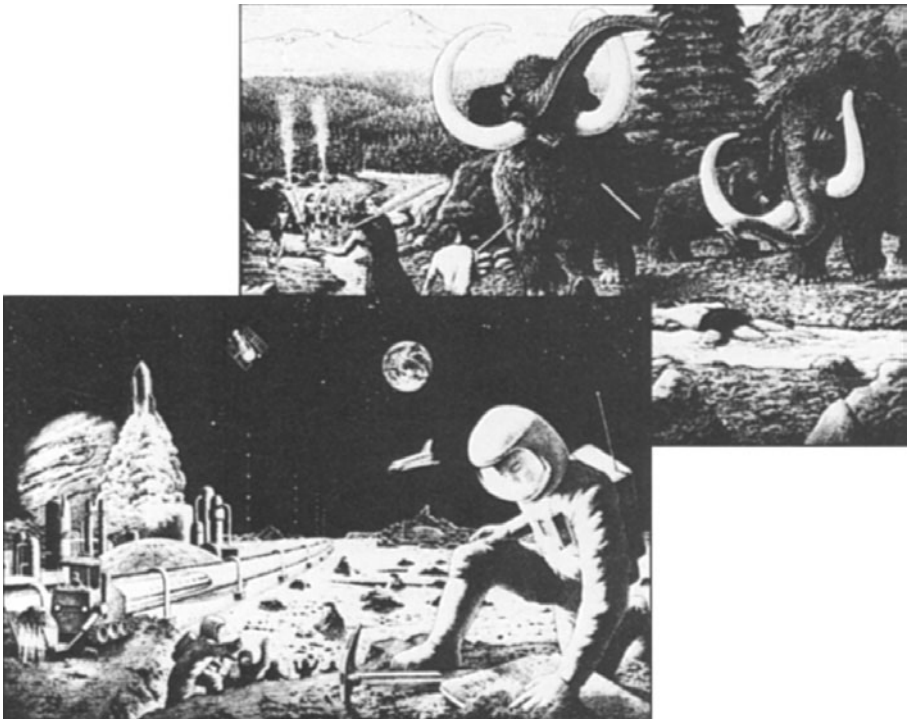


FIGURA 6.3 Muestra de imágenes de la Prueba de Lenguaje Escrito-3.
El examinado construye una historia acerca de cada una de las series de imágenes como estas dos.
(Reproducidas con autorización de pro.ed, Inc.)

(ensayo). Los estudiantes de licenciatura y de posgrado cuya lengua nativa no es el inglés pueden presentar la Prueba de Inglés Escrito (TWE) y la Prueba de Inglés Hablado (TSE). La TWE, que es aplicada por el Servicio de Pruebas Educativas junto con el TOEFL (vea líneas abajo), requiere que los examinados escriban un ensayo de 30 minutos en inglés estándar en respuesta a una breve pregunta o tema de ensayo. La TSE, que fue diseñada para medir la habilidad de hablantes no nativos del inglés para comunicarse oralmente en ese idioma, requiere que los examinados respondan de manera oral bajo condiciones temporales a una variedad de estímulos impresos y auditivos.

Los estudiantes de países extranjeros que solicitan admisión a colegios y universidades estadounidenses y cuya lengua materna no es el inglés, por lo general presentan la Prueba de Inglés como Lengua Extranjera (TOEFL). El TOEFL, un examen de opción múltiple de tres horas aplicado por el Servicio de Pruebas Educativas (ETS), consta de tres partes: Comprensión auditiva, que mide la habilidad para entender el inglés hablado; Estructura y Expresión Escrita, que mide la habilidad para reconocer el lenguaje inapropiado para el inglés estándar escrito, y Vocabulario y Lectura de Comprensión, que mide la habilidad para entender material de lectura técnico. Los estudiantes de secundaria cuya lengua materna no es el inglés, pero que desean cursar programas educativos de tiempo completo conducidos en inglés, también pueden presentar la Prueba de Dominio del Inglés de Nivel Secundaria (SLEP) (del Educational Testing Service). Otra prueba de competencia en el idioma inglés para personas cuya lengua materna no es el inglés es la Prueba de Inglés para la Comunicación Internacional (TOEIC). La prueba TOEIC, que al igual que la TOEFL y la SLEP es diseñada y administrada por el Servicio de Pruebas Educativas (ETS), es el estándar mundial para la evaluación del inglés usado en el lugar de trabajo global.

Pruebas de idiomas extranjeros. Las pruebas de estudio de la competencia en un idioma extranjero, por lo regular constan de distintas formas para estudiantes que han completado diferentes grados de preparación en ese idioma. Ciertas pruebas reflejan el enfoque gramatical más tradicional a la enseñanza del idioma, mientras que otras enfatizan la comprensión de la comunicación hablada y escrita. Las pruebas de estudio más populares de la competencia en idiomas extranjeros son los exámenes de Ubicación Avanzada del Servicio de Pruebas Educativas, los Exámenes de Materia CLEP en francés, alemán y español, y las pruebas SAT II del Consejo Universitario en esas mismas materias. También se dispone de pruebas por separado en varios idiomas en la Praxis II: Evaluaciones de Materia para Profesores Principiantes. Aunque la mayoría de las pruebas de lenguas extranjeras está limitada a la lectura y la audición, el Centro de Lingüística Aplicada administra pruebas de la habilidad para hablar chino, hausa, hebreo, indonesio, portugués y otros idiomas.

Pruebas de ciencias sociales

Los temas en ciencias sociales, historia, economía y ciencia política generalmente se consideran en conexión con los programas de estudio de secundaria y universidad. Pero las ciencias sociales, en un sentido menos restrictivo, también se enseñan en la primaria. Entre las muchas pruebas de aprovechamiento en ciencias sociales a nivel de secundaria se encuentran los Exámenes de Colocación Avanzada del Consejo Universitario a nivel de secundaria en Economía, Gobierno y Política, Historia, Geografía Humana e Historia Mundial, y los Exámenes de Materia CLEP en Gobierno Estadounidense, Historia de Estados Unidos I y II, Principios de Macroeconomía, Principios de Microeconomía, Introducción a la psicología, Introducción a la Sociología y Civilización Occidental I y II.

Pruebas de ciencias

La enseñanza de la ciencia, al igual que la de las matemáticas, cambió de manera notoria durante las pasadas tres décadas, lo cual volvió obsoletas o inapropiadas para los programas actuales de ciencias muchas pruebas antiguas. El Estudio del Currículo de Ciencias Biológicas (BSCS) y el Comité de Estudio de Ciencias Físicas (PSSC) dieron lugar al diseño de pruebas específicas en biología y física. Otros programas integrales de examinación en otras ciencias, como las Pruebas Cooperativas de Química de la Sociedad Estadounidense de Química, también reflejan enfoques contemporáneos a la educación en ciencias. Esos enfoques ponen de relieve la enseñanza del contenido de la ciencia de modo que pueda ser utilizable e importante como para incidir en la toma de decisiones de la vida cotidiana. Con esta meta en mente, las pruebas en ciencias desarrolladas más recientemente requieren que los estudiantes descubran patrones en conjuntos de datos e interpreten los significados de esos patrones en lugar de limitarse a recordarlos. Muchas pruebas antiguas también han sido revisadas en un intento por evaluar el desempeño en un programa moderno o tradicional de ciencias.

Conforme los estudiantes progresan a través de la secundaria y la preparatoria, la instrucción en ciencia general, biología, química y física se vuelve más concentrada. Los Exámenes de Ubicación Avanzada del Consejo Universitario en biología, química, ciencia ambiental y física, los Exámenes de Materia CLEP en biología general, química general y crecimiento y desarrollo humano, y las Pruebas de Materia SAT II son útiles al evaluar el conocimiento y las habilidades de estudiantes de preparatoria en campos específicos de la ciencia. Otras pruebas de ciencias para estudiantes de preparatoria y universidad incluyen los Exámenes ACS y los Exámenes de Competencia ACT.

Pruebas para la educación superior y las profesiones

Muchas instituciones de educación superior permiten que los estudiantes ganen créditos por cursos universitarios al obtener calificaciones aceptables en pruebas estandarizadas de aprovechamiento como las aplicadas por el Programa de Ubicación Avanzada del Consejo Universitario (APP), el Programa de Exámenes de Nivel Universitario (CLEP) y el Programa de Exámenes de Competencia ACT. Además, colegios, universidades y escuelas profesionales utilizan las calificaciones en las pruebas estandarizadas de aprovechamiento como criterio para la selección de estudiantes. Esas pruebas, por lo general, son *restringidas* o *seguras* en el sentido de que sólo se venden o rentan a ciertas organizaciones para su aplicación relacionada con programas educativos específicos.

Un conjunto de pruebas estandarizadas de aprovechamiento utilizadas en la selección de estudiantes para programas de posgrado lo constituyen las Pruebas de Materia de los Exámenes del Registro de Graduados (GRE). Esas pruebas, las cuales están disponibles en ocho áreas temáticas (bioquímica, biología celular y molecular; biología; química; ciencias de la computación; literatura en inglés; matemáticas; física, y psicología), pueden ser presentadas, junto con la Prueba General GRE, por estudiantes universitarios de último año que intenten solicitar admisión a la escuela de posgrado.

Otros ejemplos de pruebas estandarizadas utilizadas con propósitos de admisión a escuelas de posgrado o profesionales son la Prueba de Admisión de Administración de Graduados (GMAT), la Prueba de Admisión a la Facultad de Leyes (LSAT), la Prueba de Admisión a la Facultad de Medicina (MCAT) y las Pruebas de Aprovechamiento en Enfermería NLN. La certificación o licencia como abogado, médico, contador público, enfermera registrada, profesor o profesional en algunos otros campos también depende de aprobar una serie de pruebas de aprovechamiento (exámenes de consejo, exámenes de la barra de abogados) en el campo particular.

En Estados Unidos, 70% de los estados utiliza la Serie Praxis: Evaluaciones Profesionales para los Maestros Principiantes, como parte de su proceso para otorgar licencias a los maestros principiantes. Consta de tres partes: Praxis I: Evaluación de Habilidades Académicas, para medir las habilidades de lectura, escritura y matemáticas que son vitales para todos los candidatos a ser maestros; Praxis II: Evaluación de Materia, para medir el conocimiento que tienen los candidatos a maestros de las materias que van a impartir, y Praxis III: Evaluaciones del Desempeño en el Salón de Clases, para evaluar el desempeño del maestro principiante en el aula. Praxis I se presenta al ingresar al programa de entrenamiento de maestros, Praxis II se presenta al graduarse de la universidad e ingresar en la profesión, y Praxis III se presenta en el primer año de enseñanza.

Pruebas para administración y oficios

La administración es una materia escolar en sí misma, y las pruebas de educación en administración están diseñadas para evaluar el conocimiento que una persona tiene de la materia. Además de evaluar el grado de logro en una materia escolar, las pruebas de aprovechamiento se utilizan en la administración y la industria con propósitos de selección, colocación y promoción. Es posible que las medidas más populares sean las pruebas de competencia en mecanografía, archivo, procesamiento de palabras, cómputo y otras habilidades de oficina. Algunos ejemplos de pruebas en esta categoría son las pruebas de Mecanografía 5 y las Pruebas de Habilidades de Oficina (de NCS London House) (vea la figura 6.4).

Las pruebas de conocimiento y habilidad en un oficio (*pruebas de oficio*) se utilizan ampliamente con propósitos de selección de empleados, colocación y otorgamiento de licencia profesional. Una prueba de oficio puede consistir en una serie de preguntas que deben responderse de manera oral o escrita, o puede ser una tarea de muestra de trabajo que requiera la demostración de una habilidad en particular. Algunos ejemplos de pruebas de oficios, o de competencia ocupacional, son las proporcionadas por el programa de Desarrollo de Recursos Humanos del Servicio de Pruebas Educativas (Chauncey). Este programa ha sido responsable del desarrollo de docenas de pruebas ocupacionales o de oficios, incluyendo exámenes de competencia para certificación o licencia como inspector de código de construcción, administrador de bases de datos, planificador financiero, funcionario del servicio exterior, arquitecto paisajista, enfermera, asistente de enfermera, farmacéutico, ingeniero en plomería, podólogo, entrenador profesional de desarrollo y contador público. Por ejemplo, en la prueba para funcionario del servicio exterior, hay un "día de evaluación" en el cual se evalúa la habilidad del candidato para tomar acciones apropiadas en cada uno de un conjunto de informes y otras comunicaciones del tipo que suele encontrarse en la bandeja de un ejecutivo, así como la capacidad para manejar una entrevista de negociación de un grupo sin líder. Es obvio que esos tipos de tareas van más allá del dominio de las pruebas de habilidad y entran en el campo de la evaluación de las actitudes y la personalidad.

RESUMEN

Se administran más pruebas de aprovechamiento —al nivel de conocimiento, habilidad o logro en un área de esfuerzo— que todos los otros tipos de pruebas combinados. En el siglo pasado adquirieron cada vez más popularidad los exámenes escritos de aprovechamiento educativo, en especial los del tipo objetivo. Las pruebas objetivas pueden medir no sólo el conocimiento de hechos, sino también la comprensión y el pensamiento de orden superior. Sin embargo, se les ha criticado por alentar habilidades pobres en la composición escrita.

Reactivo muestra Mecanografía

A la persona extraviada, Ramona Woodstock, 526 Vine, se le había dicho que regresara a casa, después de visitar a Mary Lyne, no después de las 23:00 horas. Se hizo contacto con la familia a las 02:00 y la persona extraviada no había regresado a casa.

Reactivo muestra Llenado de formas

A las 8:30 am del 15 de octubre de 1977, Today's Sound Center reportó un robo en su local de 3907 Palm Ave., Wista, California. Teléfono 689-7734. Se reportó la pérdida de cuatro reproductores de cinta, dos amplificadores y dos cajas de cintas sin grabar. La puerta trasera fue forzada para poder entrar. Es posible que se haya intentado provocar un incendio en la tienda vecina para alejar la sospecha de robo. Denuncia número 789A3.

CIUDAD DE WISTA
DEPARTAMENTO DE POLICÍA

DENUNCIA Núm. *789A3*

FECHA *15 DE OCTUBRE DE 1977* HORA *8:30 AM*

NOMBRE DE LA VÍCTIMA (RAZÓN SOCIAL SI ES UNA EMPRESA): *TODAY'S SOUND CENTER*

LUGAR DE LOS HECHOS *3907 PALM AVE., WISTA* TELÉFONO *689-77-34*

PÉRDIDAS *4 REPRODUCTORES DE CINTA, 2 AMPLIFICADORES, 2 CAJAS DE CINTAS SIN GRABAR.*

Reactivo muestra Archivo

Busque el reactivo en la columna "Para archivar" y encuentre el número que debe tener este nuevo reactivo en la columna "Archivo existente". Marque con una X ese número en el renglón de números que aparecen en un círculo y están a la derecha. Si no hay número para su elección, ponga una X en el círculo en *blanco*.

Archivo existente	*Para archivar*	
1. Philip Jenkins		
2. J. C. Kile	A. B. Reynolds	① ② ③ ⊗ ○
3. Thomas Morris Company		
4. Paulson Company, Inc.	John Jones	② ③ ④ ⑤ ⊗
5. Sally White		

Reactivo Codificación

En esta prueba se le darán listas de códigos similares a la siguiente:

34 hombre
21 mujer
M adulto
U adolescente
Z niño

Debajo de las listas de códigos encontrará una lista de reactivos. Cada reactivo está seguido por círculos que contienen cinco códigos posibles. Su tarea es encontrar la combinación de códigos correcta para el reactivo y marcar con una X el círculo apropiado. Observe los siguientes ejemplos. Se ha colocado una X en la respuesta para el ejemplo 1. ¿Qué marcaría usted para el ejemplo 2?

Ejemplos:

1.	mujer adulta	(34U)	(34M) ⊗	(86Z)	(21M)	(21U)
2.	niño hombre	(21Z)	(34Z)	(34U)	(21U)	(34M)

FIGURA 6.4 Reactivos de muestra de la Prueba de Habilidades de Oficina.

Las pruebas estandarizadas de aprovechamiento reflejan objetivos educativos generales, mientras que es más probable que las pruebas elaboradas por el maestro reflejen las metas de un maestro o un sistema escolar en particular. Los resultados de las pruebas estandarizadas de aprovechamiento se utilizan para evaluar a los estudiantes con los propósitos de asignación de calificaciones, promoción, ubicación, diagnóstico de dificultades de aprendizaje, determinación de la preparación para aprender y la evaluación de los programas de estudio y la efectividad de la enseñanza (responsabilidad).

De manera tradicional, las pruebas educativas han sido sumatorias y con referencia a normas. El énfasis más reciente en la evaluación formativa, en la cual las pruebas son una parte integral del proceso instruccional, y en las pruebas con referencia a criterio es un indicador de los papeles cambiantes de las pruebas de aprovechamiento educativo. También es de importancia el uso de pruebas en la planeación y evaluación educativa a gran escala, como en la *Evaluación Nacional del Progreso Educativo.*

Cuatro tipos de pruebas de aprovechamiento son: pruebas de estudio de una materia, baterías de pruebas de estudio, pruebas de diagnóstico y pruebas de pronóstico. Las pruebas de estudio proporcionan una valoración global del aprovechamiento en una materia, mientras que las de diagnóstico analizan las fortalezas y debilidades específicas de una persona en una materia particular. Las pruebas de preparación, aptitud y otras pruebas de pronóstico intentan alentar el aprovechamiento determinando la habilidad de una persona para aprender cierto material.

Las fuentes de información relativas a las pruebas de aprovechamiento incluyen catálogos de los editores, reseñas en revistas profesionales, *Tests in Print, The Mental Measurements Yearbooks, Tests* y *Test Critiques*, grupos de muestras de pruebas y varios sitios Web (vea el apéndice C).

La confiabilidad de la mayoría de las pruebas de aprovechamiento, determinada por procedimientos de test-retest y formas paralelas, por lo general es de .80 o .90. La evidencia de la validez de contenido suele ser de mayor interés que otros tipos de validez al evaluar las pruebas de aprovechamiento educativo.

Se dispone comercialmente de varias baterías de pruebas de aprovechamiento de niveles múltiples. Esas baterías suelen aplicarse en las escuelas de primaria y secundaria. También se aplican ampliamente pruebas de una materia en lectura, matemáticas, ciencia, ciencias sociales, inglés, lenguas extranjeras y en otras áreas. Las pruebas de estudio de lectura por lo general miden el conocimiento del vocabulario, así como la velocidad y el nivel de comprensión.

Las pruebas de diagnóstico, que están diseñadas para evaluar fortalezas y debilidades específicas en una materia particular, se encuentran en lectura, aritmética y ortografía. También se dispone de varias pruebas de pronóstico en lectura (pruebas de preparación para la lectura), matemáticas y lenguaje (pruebas de aptitud para el lenguaje).

Se dispone de pruebas de aprovechamiento en ciencias sociales (historia, economía, ciencia política) y ciencias naturales (ciencia general, biología, química, física) para una amplia gama de grados y tipos diferentes de planes de estudio. También se usan de manera extensa pruebas de admisión a escuelas de enfermería (NTE), medicina (MCAT), leyes (LSAT), administración (GMAT), enseñanza (Praxis) y otros programas profesionales, y para determinar la competencia en varias ocupaciones de administración y oficios.

PREGUNTAS Y ACTIVIDADES

1. Compare las pruebas estandarizadas de aprovechamiento con las pruebas elaboradas por el maestro, mencionando los méritos y las desventajas de cada una.

2. ¿Qué es responsabilidad en educación? ¿Cómo se relaciona la responsabilidad con el contrato de desempeño? Mencione argumentos que apoyen y otros que se opongan al contrato de desempeño en las escuelas.

3. ¿En qué difiere la evaluación formativa de la evaluación sumatoria? ¿Cómo se contraponen o se complementan entre sí los dos enfoques hacia la evaluación? ¿De qué manera se relaciona la evaluación formativa con la medición con referencia a criterio?

4. Distinga entre medición con referencia a normas y medición con referencia a criterio. ¿Cuáles son las ventajas y desventajas de cada una?

5. Compare los propósitos y el diseño de las pruebas de estudio, de diagnóstico y de pronóstico.

6. Compare las pruebas donde hay mucho en juego con las pruebas donde hay poco en juego, incluyendo los tipos de prueba y las decisiones tomadas con cada una.

7. ¿En qué niveles y para qué propósitos son más válidas y útiles las pruebas estandarizadas de aprovechamiento?

8. ¿Cuáles son las ventajas y las desventajas de aplicar una batería de pruebas de aprovechamiento en lugar de una serie de pruebas sencillas de materia?

9. La mayoría de los departamentos de psicología y educación mantienen en sus archivos muestras de pruebas estandarizadas de aprovechamiento, que incluyen los folletos de la prueba, hojas de respuestas, claves de calificación, manuales y posiblemente otros materiales interpretativos. Seleccione una de esas pruebas para revisión, utilizando un perfil como el que aparece líneas abajo. Siempre que sea posible, usted debe llenar este perfil con la información obtenida al leer el manual de la prueba y examinar ésta. Espere hasta que haya completado su propia revisión antes de consultar revisiones publicadas de la prueba en *The Mental Measurements Yearbooks, Tests Critiques* u otras fuentes.

PERFIL DE REVISIÓN DE UNA PRUEBA

Contenido. Mencione el título, autor(es), editor, fecha y lugar de la publicación, formas disponibles, tipo de prueba y costo. Haga una breve descripción de las secciones de la prueba, de los tipos de reactivos que la componen y de las operaciones mentales o características que supuestamente mide. Indique cómo se seleccionaron los reactivos de la prueba y si el procedimiento de elaboración y/o la teoría en que está basada se describen con claridad en el manual.

Aplicación y calificación. Describa cualquier instrucción especial, si la prueba tiene límites de tiempo y, de ser así, cuáles son esos límites. Proporcione detalles concernientes a la calificación: como un todo, por secciones o partes y cosas similares. Indique si las instrucciones para la aplicación y la calificación son claras.

Normas. Describa el grupo o grupos (características demográficas, tamaño y cosas similares) en el o los que se estandarizó la prueba y cómo se seleccionaron las muestras (sistemática, estratificada al azar, por grupos, o de otra manera). ¿Qué tipos de normas se presentan en el manual de la prueba o en los complementos técnicos? ¿Parece ser adecuada la estandarización para los usos recomendados de la prueba?

Confiabilidad. Describa los tipos de información de confiabilidad presentados en el manual (consistencia interna, formas paralelas, test-retest, etcétera). ¿La naturaleza y los tamaños de las muestras de las que se reporta la información de confiabilidad son adecuados con respecto a los usos declarados de la prueba?

Validez. Resuma la información disponible sobre la validez (de contenido, predictiva, concurrente, de constructo) de la prueba incluida en el manual. ¿Es satisfactoria la información sobre la validez en términos de los propósitos declarados de la prueba?

Comentarios de resumen. Prepare un resumen del diseño y el contenido de la prueba y redacte un breve comentario sobre lo adecuado de ésta como medida de lo que fue diseñada para medir. ¿Proporciona el manual descripciones satisfactorias de diseño, contenido, normas, confiabilidad y validez de la prueba? ¿Qué otra información y/o datos se necesitan para mejorar la prueba y sus usos?

CAPÍTULO SIETE

TESTS DE INTELIGENCIA

Durante los inicios del siglo XX, una gran cantidad de aspirantes a psicólogos descubrieron que aplicando pruebas de inteligencia podían ganarse la vida en su profesión con algo distinto a la docencia y la investigación. Por ello, las pruebas de inteligencia en ocasiones han sido llamadas "el pan y la mantequilla de la psicología". En la actualidad las *pruebas de Binet* ya no son la única ocupación de los especialistas en psicología aplicada, pero la evaluación de aptitudes cognoscitivas todavía forma parte de las actividades de los psicólogos en los ámbitos clínicos, educativos y empresariales.

HISTORIA, DEFINICIONES Y TEORÍAS

El término *inteligencia*, común ahora en el vocabulario de la mayoría de las personas, era casi desconocido en el habla cotidiana de hace un siglo. Durante la última parte del siglo XIX, muchos académicos y científicos fueron atraídos por la teoría de Charles Darwin de que las diferencias entre las especies evolucionaban mediante selección natural. Dos de estos estudiosos, el filósofo Herbert Spencer y Francis Galton, el científico caballero primo de Charles Darwin, se interesaron por las diferencias dentro de las especies en cuanto a características mentales y comportamiento. Ambos, junto con sus seguidores, sostenían que entre los seres humanos existe un grado innato de habilidad mental general, a la que se refirieron como *inteligencia*.

A diferencia de Spencer, Galton no se contentaba simplemente con especular y discutir sobre la naturaleza de la inteligencia. Intentando demostrar que la inteligencia tiene una base hereditaria, estudió árboles genealógicos y diseñó varias pruebas de discriminación sensorial y tiempo de reacción para medir sus componentes. Éstas y otras pruebas sensoriomotrices (velocidad de movimiento, fuerza muscular, sensibilidad al dolor, discriminación de peso y otras similares) fueron estudiadas ampliamente por el psicólogo estadounidense J. McKeen Cattell. Desafortunadamente, las pruebas resultaron relativamente inútiles para predecir el desempeño en tareas escolares y otras actividades que supuestamente requieren de inteligencia.

El enfoque del psicólogo francés Alfred Binet fue radicalmente distinto al procedimiento analítico de tratar de medir los componentes de la inteligencia. Binet sostenía que la inteligencia se manifiesta en el desempeño en diversas tareas y que podía medirse mediante respuestas a una muestra de dichas tareas. Debido a que el trabajo de Binet al diseñar las primeras pruebas de inteligencia con éxito fue motivado por el problema de identificar niños con retraso mental en el sistema escolar de París, es natural que la muestra de pruebas seleccionada por él estuviera plagada de tareas de tipo escolar.

En 1905 Binet y su socio, el doctor Théodore Simon, publicaron su primera serie de pruebas de inteligencia, 30 pruebas breves ordenadas desde la más sencilla hasta la más difícil. Al

proseguir su trabajo, publicaron en 1908 una escala modificada Binet-Simon que consistía en 58 tareas dispuestas por niveles de edad de 3 a 13 años. Las tareas se agruparon por edad cronológica de acuerdo con lo que había indicado la investigación que podrían realizar los niños normales de una edad determinada. La edad mental (MA [EM]) de un niño se establecía por la cantidad de subpruebas aprobadas en cada nivel, y una edad mental notablemente inferior a la edad mental del niño se consideraba indicativa de retraso mental. En 1911 se publicó una última versión modificada de la escala (tabla 7.1), pero después de la muerte prematura de Binet en ese mismo año, la escena de los posteriores desarrollos en cuanto a pruebas de inteligencia se mudó a Estados Unidos y Gran Bretaña.

Definición de la inteligencia

Desde que Binet y Simon produjeron las primeras pruebas prácticas de inteligencia, los psicólogos han intentado formular una definición viable del concepto. La explicación de Binet destacaba el juicio, el entendimiento y el razonamiento. Otras definiciones describían la inteligencia como la habilidad de pensar en forma abstracta, la habilidad de aprender o la habilidad de adaptarse al medio ambiente. Sin embargo, todas estas definiciones fueron criticadas por una u otra razón. La habilidad obviamente es necesaria para la sobrevivencia, pero resulta una definición de la inteligencia demasiado amplia. Por otra parte, la definición de inteligencia de Lewis Terman como la habilidad de tener pensamiento abstracto es demasiado estrecha. La habilidad para el pensamiento abstracto es un aspecto importante de la inteligencia, pero ciertamente no es el único. Por último, la concepción popular de inteligencia como la habilidad de aprender es inadecuada si se aceptan las pruebas de inteligencia como medida de ésta. Los aciertos en tales pruebas no están correlacionados en alto grado con el ritmo o la velocidad de aprender cosas nuevas, aunque sí están más relacionados con el nivel o la cantidad de aprendizaje.

Más que intentar formular una definición universalmente aceptable de la inteligencia, algunos psicólogos han sugerido que podría ser mejor abandonar el término por completo. Si se requiere un término alternativo, tal vez sería preferible utilizar *habilidad mental general,* o *habilidad académica.* Los dos últimos términos son un reconocimiento al hecho de que las pruebas de inteligencia tradicional son sobre todo predictores del éxito en el trabajo escolar. Sin importar lo intensa que pueda ser la oposición al término *inteligencia*, es ciertamente menos fuerte que la oposición al coeficiente intelectual (CI). Debido a la controversia existente sobre el CI y a la implicación de que es una medida fija de habilidad cognoscitiva, ciertos psicólogos que han dedicado gran parte de sus vidas profesionales al estudio de la inteligencia han expresado una disposición a abandonar por completo el término CI (Vernon, 1979).

No todos los instrumentos examinados en este capítulo tienen la etiqueta específica de *prueba de inteligencia*; más bien se han propuesto como medidas de habilidad mental *general*. En este sentido, deben distinguirse de las medidas de *habilidades especiales* consideradas en el capítulo 10. Sin embargo, no está clara la distinción entre pruebas de habilidad mental general (inteligencia) y pruebas de habilidades especiales, y ciertas pruebas de habilidad académica analizadas en este capítulo podrían corresponder igualmente bien al capítulo 10.

Teorías de la inteligencia

Las teorías de la inteligencia, o más bien del comportamiento inteligente, se han basado en modelos psicométricos de desarrollo y procesamiento de información. Los primeros dos tipos de teorías son enfoques tradicionales, la tercera teoría es de origen más reciente.

TABLA 7.1 Las cincuenta y cuatro subpruebas de la Escala de Inteligencia Binet-Simon de 1911

3 años de edad
Señala sus ojos, nariz y manos.
Repite dos dígitos.
Enumera objetos de una imagen.
Dice su apellido.
Repite una oración de seis sílabas.

4 años de edad
Dice su sexo.
Nombra llave, cuchillo, dinero.
Repite tres dígitos.
Compara dos líneas.

5 años de edad
Compara dos pesos.
Copia un cuadrado.
Repite una oración de diez sílabas.
Cuenta cuatro centavos.
Une las mitades de un rectángulo dividido.

6 años de edad
Distingue entre mañana y tarde.
Define palabras familiares en términos de uso.
Copia un rombo.
Cuenta 13 monedas.
Distingue dibujos de rostros feos y hermosos.

7 años de edad
Muestra su mano derecha y su oreja izquierda.
Describe un dibujo.
Ejecuta tres órdenes dadas simultáneamente.
Cuenta el valor de seis centavos, tres de los cuales son dobles.
Nombra cuatro colores principales.

8 años de edad
Compara dos objetos de memoria.
Cuenta de 20 a cero.
Señala omisiones en dibujos.
Da el día y la fecha.
Repite cinco dígitos.

9 años de edad
Da cambio de 20 centavos.
Define palabras familiares en términos superiores al uso.
Reconoce todas las monedas (nueve).
Nombra los meses del año en orden.
Contesta o comprende "preguntas fáciles".

10 años de edad
Ordena cinco bloques por peso.
Copia dos dibujos de memoria.
Critica afirmaciones absurdas.
Contesta o comprende "preguntas difíciles".
Usa tres palabras dadas en no más de dos enunciados.

12 años de edad
Resiste sugerencias sobre la extensión de líneas.
Compone una oración con tres palabras dadas.
Menciona 60 palabras en 3 minutos.
Define tres palabras abstractas.
Descubre el sentido de una oración desordenada.

15 años de edad
Repite siete dígitos.
Encuentra tres rimas para una palabra dada en un minuto.
Repite una oración de 26 sílabas.
Interpreta imágenes.
Interpreta hechos dados.

Adulto
Resuelve el test del papel cortado.
Reacomoda un triángulo en la imaginación.
Menciona diferencias entre pares de términos abstractos.
Da tres diferencias entre un presidente y un rey.
Encuentra la idea principal en un párrafo que ha leído.

Teorías psicométricas. El método psicométrico, que ha dado origen a muchas pruebas de inteligencia y diversos métodos estadísticos para analizar las calificaciones de estas pruebas, se centra en las diferencias individuales en cuanto a habilidades cognoscitivas y en la búsqueda de las causas de estas diferencias. Entre las teorías o modelos de habilidades cognoscitivas basadas en el método psicométrico y originadas sobre todo de los resultados del análisis factorial (vea apéndice A), figuran la teoría bifactorial de Spearman (1927) (que consiste en un factor general

más varios factores específicos para cada prueba), la teoría multifactorial de siete habilidades mentales básicas de Thurstone (Ekstrom, French y Harman, 1979), el modelo de estructura del intelecto de Guilford (1985) y el modelo jerárquico de Vernon (1960). El modelo de Vernon consiste en un factor general en el primer nivel, factores verbales-educacionales y práctico-mecánico-espaciales en el segundo nivel, y varios factores de un grupo menor en un tercer nivel (vea la figura 7.1). La teoría de Cattell (1963) de dos tipos de inteligencia, fluida y cristalizada, también se basa en los resultados del análisis factorial y se relaciona con la distinción de Hebb (1949) entre Inteligencia A e Inteligencia B.

Teorías sobre el desarrollo. Las teorías sobre el desarrollo de las habilidades cognoscitivas que provienen de la investigación sobre psicología del desarrollo humano, subrayan la uniformidad o las similitudes interindividuales en la evolución cognoscitiva más que las diferencias individuales. Un ejemplo primordial es la idea de Piaget de que la cognición se desarrolla a partir de las acciones de asimilación y acomodamiento en el mundo exterior. La *asimilación* consiste en ajustar las nuevas experiencias en las estructuras cognoscitivas preexistentes (esquemas *schemata*); el *acomodamiento* es la modificación de estos *schemata* como resultado de la experiencia. Al interactuar con el ambiente, un niño en crecimiento crea *schemata* de modo que funcionen como mapas explicativos y guías para el comportamiento. De acuerdo con Piaget, por lo regular los niños se desarrollan intelectualmente a través de una serie de etapas progresivas: sensoriomotriz (del nacimiento a 2 años de edad), preoperativa (de 2 a 7 años de edad), operativa concreta (de 7 a 11 años de edad) y operativa formal (de 11 a 15 años de edad). Piaget pensaba que el aumento de la inteligencia se detenía a la edad aproximada de 15 años, pero varios investigadores han objetado esta afirmación.

Teorías sobre el procesamiento de información. Las teorías sobre procesamiento de información, o modelos de resolución de problemas y razonamiento, se ocupan de identificar los procesos cognoscitivos u operacionales mediante los cuales el cerebro maneja la información. La investigación sobre atención y velocidad de procesamiento ha recibido un énfasis particular desde una perspectiva de procesamiento de información. Resultan ilustrativas de las teorías de procesamiento de información las teorías triárquicas o de proceso componencial de Sternberg

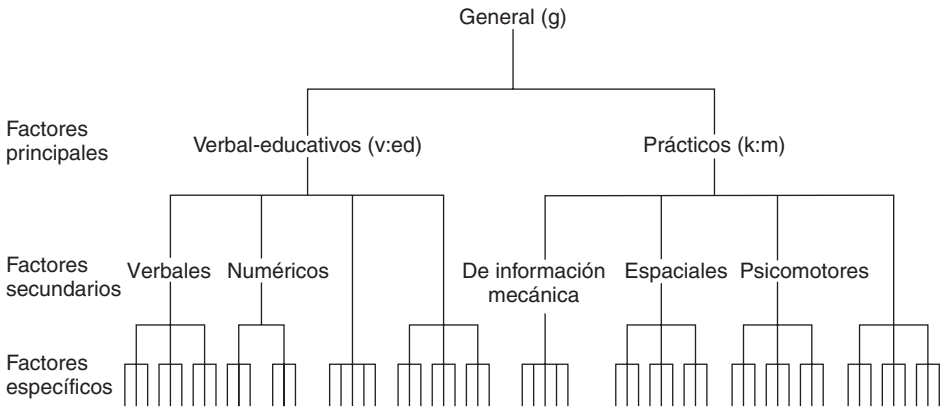


FIGURA 7.1 Modelo jerárquico de Vernon de las habilidades intelectuales.

(De acuerdo con Vernon, 1960, p. 22. Reproducida con autorización de la editorial Routledge.)

(1982), la teoría de inteligencias múltiples de Gardner (1983), y el modelo PASS de Das, Naglieri y Kirby (1994).

En un principio, Sternberg (1982) formuló la hipótesis de que existen cinco clases de *procesos componenciales* mediante los cuales el cerebro opera sobre la información y resuelve problemas, a saber: metacomponentes, componentes de desempeño, componentes de adquisición, componentes de retención y componentes de transferencia. Entre los diversos componentes de estas cinco clases, la codificación y la comparación son especialmente críticas para lograr una resolución efectiva de los problemas. En una extensión de su teoría de procesos componenciales, Sternberg (1985, 1986) propuso una *teoría triárquica* que incluye tres subteorías: componencial, experiencial y contextual. La subteoría componencial consiste en metacomponentes, componentes de desempeño y componentes de adquisición de conocimiento. La subteoría experiencial se ocupa de la de formular nuevas ideas combinando factores o información aparentemente no relacionados. La subteoría contextual aborda la de adaptarse a condiciones ambientales cambiantes y conformar el entorno de tal modo que nuestras ventajas se incrementen y nuestras desventajas se compensen. En una modificación posterior de su teoría, Sternberg (1988) propuso el concepto de *autocontrol mental*, que representa un intento por combinar el concepto de inteligencia con el de personalidad. Las maneras en que los tres tipos de inteligencia delineados por la teoría triárquica —componencial, experiencial y contextual— se ponen en práctica en la resolución de los problemas cotidianos, se caracterizan como *estilos intelectuales*. La efectividad de un estilo intelectual en particular depende de la medida en que se ajuste a la capacidad intelectual de la persona, su estilo preferido y el problema inmediato por resolver.

De acuerdo con la teoría de Gardner de inteligencias múltiples, la cognición y el procesamiento de información en los humanos implica el despliegue de varios sistemas simbólicos que son formas características de percepción, memoria y aprendizaje. Gardner propuso que hay siete formas de inteligencia: lingüística, lógico-matemática, espacial, musical, kinestésica corporal, y dos formas de inteligencia personal (intrapersonal e interpersonal). Sostuvo que sólo las primeras tres formas se miden mediante tests de inteligencia convencionales, y que la cultura occidental ha puesto demasiado énfasis en la primera de éstas, la lingüística. Sin embargo, Gardner advierte que las otras dos formas de inteligencia (lógico-matemática y espacial) son más valiosas en muchas sociedades y circunstancias.

El modelo de inteligencia PASS (planeación, atención, procesamiento simultáneo, procesamiento sucesivo) se basa en la teoría de Aleksandr Luria de que el cerebro humano está funcionalmente dividido en tres unidades. La *primera unidad funcional*, que se asocia con el tallo cerebral superior y el sistema límbico, es responsable de la estimulación y la atención. La *segunda unidad funcional* está asociada con las regiones posteriores de los hemisferios cerebrales, incluyendo las áreas visual (occipital), auditiva (temporal) y sensorial general (parietal); es responsable de la recepción, el análisis y almacenamiento de la información mediante procesos de razonamiento simultáneos y sucesivos. La *tercera unidad funcional* está asociada con las partes anteriores del hemisferio cerebral, en particular con la región prefrontal; es responsable de planear, regular y verificar la actividad cognoscitiva. Para efectuar el procesamiento cognoscitivo de información, la base de conocimiento del individuo debe estar integrada con los procesos de *planeació*n (tercera unidad funcional), *atención* (primera unidad funcional), y procesos *simultáneos y sucesivos* (segunda unidad funcional) como lo requiere una tarea en particular. El resultado de semejante proceso cognoscitivo incluye hablar, escribir u otras actividades motoras (Das, Naglieri y Kirby, 1994).

A pesar de éstos y otros intentos interesantes y valerosos, ningún método teórico ha logrado proporcionar una explicación totalmente satisfactoria sobre cómo la inteligencia se desarro-

lla y cambia, las causas de las diferencias individuales en la inteligencia, o los procesos cognoscitivos y fisiológicos específicos que son responsables de la actividad intelectual. Al parecer, todas las corrientes actuales son correctas en cierta medida, pero de seguro ninguna proporciona una explicación completa, empíricamente verificada, sobre la estructura y el funcionamiento cognoscitivo. Por el momento, parece que las teorías sobre procesamiento de información ofrecen la mejor oportunidad de lograr una concepción lógica y con base empírica de las habilidades cognoscitivas, pero la situación podría cambiar al avanzar las investigaciones. De cualquier modo, algo es cierto: surgirán otras teorías sobre la inteligencia, y su valor se determinará por su eficacia para predecir y explicar el aprendizaje y el pensamiento humanos.

Aplicaciones de las evaluaciones de inteligencia

En contraste con otras definiciones más teóricas, las definiciones operativas de la inteligencia se centran en su medición y en las aplicaciones relacionadas. Tal vez la más operativa de dichas definiciones fue la sugerida por E. G. Boring, quien propuso definir la inteligencia como "aquello que se mide por medio de un test de inteligencia". Lo que sea que midan los tests de inteligencia, estas pruebas se han usado para varios fines prácticos, incluyendo (1) el diagnóstico de la habilidad mental alta y baja y la ubicación de los retrasados mentales o los superdotados en programas o clases especiales; (2) la selección (sondeo), colocación y clasificación de estudiantes en instituciones de educación superior, empleados en organizaciones de negocios o industriales y personal en dependencias militares y gubernamentales; (3) la determinación y el diagnóstico de discapacidades relacionadas con el trabajo por demandas de seguros; (4) la asesoría y rehabilitación vocacional y educativa; (5) el psicodiagnóstico de niños y adultos en contextos clínicos o psiquiátricos; (6) la evaluación de la efectividad de tratamientos psicológicos e intervenciones en el medio ambiente, y (7) los estudios sobre habilidades cognoscitivas y personalidad.

Pruebas individuales colectivas

A pesar del objetivo común de medir una habilidad unitaria, los formatos de todos los tests de inteligencia general no son idénticos. En algunos hay reactivos de distintos tipos mezclados o alternados, y aumenta su dificultad a lo largo de la prueba. Los reactivos de otros tests de inteligencia se agrupan como conjuntos de subpruebas programadas en forma separada.

La forma más común de clasificar las pruebas de inteligencia es mediante la dicotomía *individual* versus *colectiva o de en grupo*. Los tests de inteligencia individual, que se aplican a una persona a la vez, tienen un enfoque algo distinto que los tests de inteligencia colectiva, los cuales pueden administrarse a muchas personas simultáneamente. El énfasis de las pruebas individuales es más global u holístico: su principal función es evaluar una habilidad cognoscitiva general. Por otra parte, el enfoque del test colectivo tiende a ser más reducido: a predecir el desempeño académico o laboral. Además, administrar un test de inteligencia individual suele ser más laborioso que administrar una prueba . Una ventaja de las pruebas individuales es que los examinadores pueden prestar más atención a los sujetos de examen. El enfoque del examinado a la prueba y otros comportamientos —angustia, confianza, estrategias para resolver problemas, frustraciones, distracción y aspectos similares— pueden observarse más de cerca cuando se examina a una persona a la vez, y el desempeño puede estimularse y recompensarse en forma más efectiva. Asimismo, las calificaciones de pruebas individuales no dependen tanto de la capacidad de lectura como las calificaciones de pruebas aplicadas colectivamente.

La mayor economía de administrar una prueba en grupo en ciertas situaciones ocasiona que se administren más pruebas en grupo que individuales. Además, a pesar de lo que en ocasiones han sostenido los defensores de las pruebas individuales, ciertas pruebas de inteligencia aplicadas en forma colectiva grupal pueden incluso tener mayores coeficientes de validez que sus contrapartes individuales.

Los tests de inteligencia colectivos grupales se usan con mayor frecuencia para una selección inicial en situaciones educativas y laborales, que es seguida por una evaluación individual cuando el examinado obtiene una calificación deficiente en una prueba colectiva y/o se requiere más información sobre sus cualidades y fallas cognoscitivas. También es más probable que los tests de inteligencia individuales se usen en clínicas, hospitales y otros sitios donde se realizan diagnósticos clínicos. En dichos lugares las pruebas sirven no sólo como medidas de la habilidad mental general, sino también como medio de comprender más a fondo el funcionamiento de la personalidad y las discapacidades cognoscitivas específicas.

TESTS DE INTELIGENCIA INDIVIDUALES

Los instrumentos que provienen del trabajo de Lewis Terman y David Wechsler han sido las pruebas de inteligencia individuales más comunes. Con el paso del tiempo, estos tests se han usado para evaluar las habilidades intelectuales de niños y adultos en muchos contextos diferentes. Otras pruebas individuales, algunas de las cuales constituyen variantes o extensiones de los tests de Terman y de Wechsler, se han diseñado específicamente para evaluar las habilidades mentales de niños pequeños y personas con desventajas lingüísticas y/o físicas.

Otras ediciones de la Escala de Stanford-Binet

Hubo tres traducciones y adaptaciones de la escala Binet-Simon en Estados Unidos. Una fue preparada por H. H. Goddard de la Escuela de Capacitación Vineland, otra por Frederic Kuhlmann de la Universidad de Minnesota, y una tercera por Lewis Terman de la Universidad Stanford. La más popular de estas revisiones, la Escala de Inteligencia Stanford-Binet, fue publicada por Terman en 1916.

La Escala de 1916. Al igual que las anteriores escalas de Binet-Simon, la Stanford-Binet de 1916 era una escala de edad donde las subpruebas se agrupaban en niveles de edad cronológica. Terman seleccionó reactivos de las escalas de Binet-Simon, así como reactivos totalmente nuevos que representaban una muestra amplia de las tareas que supuestamente requerían capacidades intelectuales aprovechadas. También se realizaron esfuerzos para incluir tareas que no eran tan dependientes de experiencias de aprendizaje escolares específicas.

Un criterio para incluir un reactivo en la escala de Stanford-Binet era que un porcentaje creciente de niños en niveles de edad sucesivos deberían ser capaces de responder el reactivo en forma correcta. Por algunas razones estadísticas que tienen que ver con mantener una escala de cociente de inteligencia bastante estable a través de los niveles de edad, el porcentaje de aprobados requerido se estableció más bajo en reactivos incluidos en subpruebas en niveles de años superiores que en reactivos de niveles de años inferiores. De cualquier modo, el criterio del porcentaje de aprobados sirvió como un medio objetivo de asegurarse que cada reactivo de la prueba se ubicara en un nivel de edad adecuado.

La *edad mental* (EM) y el *cociente de inteligencia* de un examinado en la escala Stanford-Binet dependían de la cantidad de subpruebas aprobadas en los niveles de edad sucesivos. El cociente de inteligencia se determinaba dividiendo la edad mental del examinado (EM), la cantidad

total de crédito de meses obtenida en la prueba, por su edad cronológica (EC) en meses y multiplicando el cociente resultante por 100. En símbolos, esta *razón de CI* se calculaba como:

$$CI = 100 \frac{MA}{CS} \quad \textbf{(7.1)}$$

Durante muchos años, la Escala de Inteligencia de Stanford-Binet funcionó como un estándar con respecto al cual se evaluaban otros tests de inteligencia. Sin embargo, tenía varias desventajas. Por ejemplo, la versión de 1916 sólo se estandarizó en 1,000 niños y 400 adultos. De acuerdo con las normas actuales, la muestra no se seleccionó con cuidado y no era representativa de la población estadounidense de la época. Otras dos desventajas fueron la inadecuación al evaluar adultos y niños muy pequeños, y la falta de una segunda forma para permitir la revaluación. Por lo tanto, en 1937, Terman y su socia, Maud Merrill, publicaron una versión revisada, actualizada y reestandarizada de la escala.

La Escala de 1937. La versión de 1937 de la Escala de Inteligencia de Stanford-Binet tenía un límite *inferior* menor *y* uno *superior* mayor que la escala de 1916, dos formas paralelas (L y M) y una mejor estandarización. La escala de 1937 fue estandarizada de manera estratificada en 100 niños, con un intervalo por cada medio año de edad, desde el año y medio hasta los cinco y medio años; 200 niños con intervalos por cada año de edad desde los 6 hasta los 14 años, y 100 niños con intervalos por cada año de edad desde los 15 hasta los 18 años. Se administró la prueba a un número igual de niñas y niños en 17 comunidades de 11 estados, pero la muestra se limitó a individuos blancos nativos, quienes, como grupo, estaban en cierta medida por encima del promedio en cuanto a situación socioeconómica. En consecuencia, la muestra no era verdaderamente representativa de toda la población de Estados Unidos.

Se usaron tres criterios para incluir un reactivo en la escala: (1) el reactivo se consideró como una medida de comportamiento inteligente; (2) el porcentaje de niños que pasaban el reactivo aumentaba con la edad cronológica, y (3) los niños que aprobaron el reactivo tenían una edad mental media superior que la de quienes fracasaron en el reactivo. Los reactivos se agruparon en intervalos de medio año (niveles) del Año II al Año V, y en intervalos de un año desde el Año VI hasta el Año XIV; también había nivel Promedio de Adultos y tres niveles Superiores de Adulto (Adulto Superior I, II y III). Cada una de las seis subpruebas por nivel desde el Año II hasta el Año V recibió un mes de crédito, y las seis subpruebas en niveles Superiores de Adulto I, II y III tuvieron 4-, 5- y 6- meses de crédito, respectivamente.

Al evaluar a un niño con la Escala Stanford-Binet, el examinador primero determinaba la *edad basal* del niño. La edad basal era el nivel de años más alto en que el niño pasaba todas las subpruebas. La evaluación continuaba entonces hasta la *edad tope*, el nivel de años inferior en que el niño fallaba en todas las pruebas. La edad mental se calculaba añadiendo a la edad basal el número de meses de crédito recibido por pasar cada subprueba hasta la edad tope. Entonces de calculaba el CI mediante la fórmula 7.1.

La Escala de 1960. La tercera edición de la Escala de Inteligencia de Stanford-Binet, publicada en 1960, consistía en una actualización de los mejores reactivos de las formas L y M. Al igual que sus predecesoras, la tercera edición se usaba para medir la inteligencia de individuos desde la edad de dos años hasta la adultez. El procedimiento para administrar la prueba era similar al de la escala de 1937, pero se introdujeron algunos cambios. Uno de éstos consistía en una subprueba alternativa en cada nivel de edad para usarla cuando alguna de las subpruebas no se

aplicaba o se aplicaba de modo incorrecto. El tiempo de la prueba también podía reducirse en ciertos casos administrando sólo cuatro subpruebas seleccionadas en lugar de seis en cada nivel de un año. Otro cambio fue la disposición para prevenir desviaciones del CI. La razón del CI, al igual que cualquier otra norma de edad, no satisfizo el requisito de igualdad de unidades de edad. Asimismo, no tenía sentido cuando se aplicaba a adultos, porque no había una respuesta satisfactoria a la pregunta sobre qué edad cronológica debía usarse como denominador de la relación MA/CA al evaluar adultos. Se han propuesto las edades de 14, 16 y 18 años como la edad en que el crecimiento mental se detiene y, por lo tanto, cualquiera de esas edades puede ser un denominador adecuado para calcular el CI. Debido a los problemas para determinar la razón del CI, se tomó la decisión de cambiar de un CI de razón a una calificación estándar escala de *desviación CI*, con una media de 100 y desviación estándar de 16. Ocasionalmente se siguió reportando la razón de CI antigua y se incluían tablas para calcularla en el manual de Stanford-Binet de 1960.

La muestra de estandarización para la Forma 1960 L-M de la Escala Stanford-Binet consistió en 4,500 niños, de entre $2^{1}/_{2}$ y 18 años de edad, que habían tomado cualquiera de las formas L o M de la Escala de 1937 entre 1951 y 1954. Tomando en cuenta la necesidad de normas actualizadas, el editor hizo adaptaciones para la prueba al ser administrada en 1972 a una muestra nacional estratificada de 2,100 niños (100 niños por cada intervalo de medio año desde los 2 hasta los $5^{1}/_{2}$ años, y por cada intervalo de un año también 100 niños, éstos de 6 a 18 años). La muestra era más representativa que las anteriores muestras normativas de la población general de Estados Unidos. Con base en la estandarización de 1972, se publicó un manual revisado para la tercera edición (Terman y Merrill, 1973). El manual incluía coeficientes de confiabilidad de test-retest de más de .90 y, como en las dos primeras ediciones, correlaciones moderadas con grados escolares y calificaciones de pruebas de aprovechamiento (.40 a .75).

Cuarta edición de la Escala Stanford-Binet

La cuarta edición de la Escala de Inteligencia Stanford-Binet (SB-IV) (por Riverside Publishing) se elaboró considerando las necesidades de psicólogos clínicos, escolares y otros psicólogos que usan la información de los tests de inteligencia. SB-IV mantuvo la continuidad histórica con las versiones anteriores de la escala, pero representó una marcada separación de sus predecesoras en cuanto a sus bases teóricas y psicométricas, su contenido y el procedimiento de administración. Al igual que muchas pruebas modernas, SB-IV fue desarrollada usando procedimientos psicométricos complejos, tales como la teoría de respuesta al ítem (escala de Rasch) y análisis de sesgo étnico. Además estaba diseñada no sólo para ayudar a identificar individuos con retraso mental o superdotados, sino también a proporcionar información diagnóstica sobre discapacidades de aprendizaje específicas. Con respecto al sesgo por sexo y etnia, se omitieron los reactivos considerados injustos o que mostraban diferencias estadísticas atípicas entre sexos o grupos étnicos.

Modelo teórico y pruebas. Como se diagrama en la figura 7.2, el modelo en que se basó la escala SB-IV consiste en una jerarquía de tres niveles con un factor de inteligencia general (*g*) en el primer nivel, tres factores amplios (habilidades cristalizadas, habilidades fluido-analíticas y memoria de corto plazo) en el segundo nivel, y tres factores (razonamientos verbal, cuantitativo y abstracto-visual) en el tercer nivel. Los factores de razonamiento cuantitativo y verbal comprenden el factor de habilidades cristalizadas en el segundo nivel, y el factor abstracto-visual en el tercer nivel comprende el factor de habilidades fluido-analíticas en el segundo nivel.

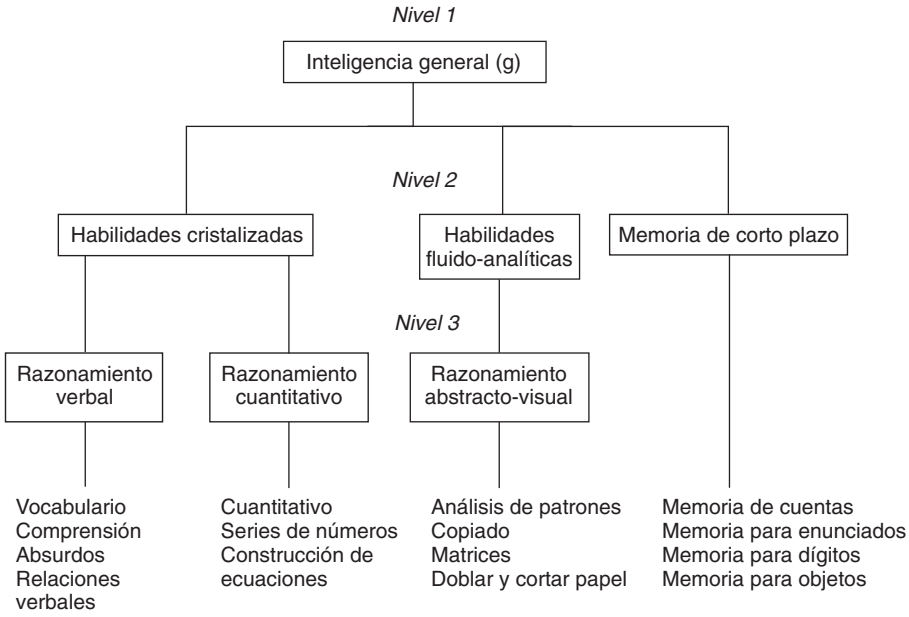


FIGURA 7.2 Modelo teórico y pruebas para la escala Stanford-Binet IV.

(Derechos Reservados 1986 por Riverside Publishing Company. Reproducido de *The Stanford-Binet Intelligence Scale*, cuarta edición, por Robert L. Thorndike, Elizabeth P. Hagen y Jerome M. Sattler, con autorización del editor.)

Al igual que sus antecesoras, la escala SB-IV fue diseñada para medir la inteligencia desde los 2 años hasta la edad adulta. Hay 15 tests: tres o cuatro tests en cada una de las tres categorías más amplias del Nivel 3 (Razonamiento Verbal, Razonamiento Cuantitativo, Razonamiento Abstracto-Visual), además de cuatro tests de Memoria de Corto Plazo (vea la figura 7.2). Cada prueba se acomoda en una serie de niveles que consisten en dos reactivos cada uno. Casi todas las pruebas incluyen reactivos de muestra para familiarizar a los examinados con el carácter de la tarea específica.

Aplicación. El tiempo de administración para toda la escala SB-IV es de aproximadamente 75 minutos, y varía de acuerdo con la edad del examinado y la cantidad de pruebas administradas. El carácter adaptativo, o de múltiples etapas, de la prueba exige administrar el Test de Ruta (Vocabulario) primero para determinar el nivel inicial en los demás tests. El nivel de entrada en el Test de Ruta se determina por la edad cronológica del examinado. La administración del Test de Ruta continúa mientras el examinado no falle en tres o cuatro reactivos en dos niveles consecutivos, el más alto de los cuales es el *nivel crítico*. El nivel inicial para las 14 pruebas restantes se establece a partir de la tabla mediante una combinación del nivel crítico del Test de Ruta y la edad cronológica del examinado (*nivel basal*) y hacia arriba hasta que falla en tres o cuatro reactivos en dos niveles consecutivos. El más alto de estos niveles es la *edad tope* del examinado para esa prueba.

Calificación. Las puntuaciones crudas en cada una de las 15 pruebas son iguales a la cantidad de reactivos aprobados. Estas puntuaciones se convierten, dentro de cada grupo de edad, en calificaciones normalizadas de *escala de edad estándar* (SAS) con una media de 50 y desviación estándar de 8. Las puntuaciones crudas en cada una de las cuatro áreas (Razonamiento Verbal, Razonamiento Abstracto-Visual, Razonamiento Cuantitativo, Memoria de Corto Plazo) son iguales a la suma de las puntuaciones crudas en las tres o cuatro pruebas que comprenden esa área. Estas puntuaciones de área se convierten en calificaciones de escala estándar (*calificaciones de área SAS*) con una media de 100 y desviación estándar de 16. Por último, una *calificación compuesta* que consiste en la suma de las cuatro puntuaciones de área se convierte a una escala de calificaciones estándar con una media de 100 y desviación estándar de 16. El rango de las calificaciones compuestas generales es de 36 a 164, que es el equivalente a un rango de calificaciones z de -4 a $+4$.

Estandarización. En Estados Unidos, la escala SB-IV fue estandarizada en 5,013 individuos de entre 2 y 23 años 11 meses de edad en 47 estados y el Distrito de Columbia. La muestra de estandarización fue estratificada por género y raza-etnia, y los estudiantes también fueron estratificados de acuerdo con la posición relativa en su clase. A pesar de los esfuerzos por seleccionar una muestra de estandarización que fuese verdaderamente representativa de la población estadounidense, la muestra contenía cantidades desproporcionadas de individuos de los niveles socioeconómicos y educativos más altos. Se intentó corregir este error al calificar las pruebas, pero el esfuerzo no fue del todo exitoso. Otros problemas son que los factores medidos por la escala no son uniformes en todos los niveles de edad y la información de confiabilidad del manual es inadecuada. Sin embargo, los coeficientes de división por mitad y de test-retest, calculados en medidas obtenidas a lo largo de un intervalo de 2 a 8 meses indican que las confiabilidades de las 15 pruebas, las cuatro áreas, y el conjunto son satisfactorias.

Las pruebas de Wechsle

Aunque las subpruebas en el nivel adulto se han incluido en la escala Stanford-Binet desde la revisión de 1937, nunca ha habido una medida muy satisfactoria de la inteligencia en adultos. Por consiguiente, en 1939 David Wechsler, un psicólogo del Hospital Bellevue en Nueva York, publicó un test de inteligencia individual diseñado específicamente para adultos. Para esta prueba, la Forma I de la Escala de Inteligencia de Wechsler-Bellevue, Wechsler añadió una segunda forma en 1947, la Forma II de la Escala de Inteligencia de Wechsler-Bellevue. Una revisión completa y reestandarización de la Forma I se publicó en 1955 como la Escala de Inteligencia para Adultos de Wechsler (WAIS). La WAIS misma fue modificada, reestandarizada y reeditada por The Psychological Corporation en 1981 como la Escala de Inteligencia para Adultos de Wechsler Revisada para evaluar la inteligencia de adultos entre 16 y 74 años de edad.

Escala de Inteligencia para Adultos de Wechsler, Revisada. Las seis subpruebas verbales (V) y cinco subpruebas de ejecución (E) de la Escala de Inteligencia para Adultos de Wechsler Revisada (WAIS-R), por orden de administración, se describen en la tabla 7.2. Las subpruebas Verbal y de Ejecución se administran alternadamente, y dentro de cada subprueba los reactivos se presentan en orden de dificultad creciente. Se requieren alrededor de 75 minutos para administrar las once subpruebas, y la administración en una subprueba en particular se descontinúa cuando el examinado falla en una cantidad específica de reactivos sucesivos.

Calificación. Las puntuaciones crudas en las once subpruebas del WAIS-R se convierten a una escala de calificación normalizada con una media de 10 y desviación estándar de 3. Entonces,

TABLA 7.2 Subpruebas de la Escala de Inteligencia para Adultos de Wechsler Revisada

Información (*V*): 33 preguntas sobre información general que deben contestarse en pocas palabras o números.

Completamiento de dibujos (E): 27 dibujos en tarjetas, cada uno con una parte faltante; el examinado tiene 20 segundos para indicar lo que hace falta en la imagen.

Serie de dígitos (*V*): 7 series de dígitos que deben recitarse hacia adelante y 7 series para repetirse hacia atrás.

Ordenamiento de dibujos (E): 10 series de tarjetas, cada una con una pequeña imagen; se pide al examinando que ordene los dibujos de cada serie de tarjetas para construir una historia coherente.

Vocabulario (*V*): se presentan 37 palabras en orden de dificultad creciente que deben definirse.

Diseño de cubos (E): 10 diseños geométricos en rojo y blanco en tarjetas y nueve bloques de los mismos colores; se solicita al examinado que copie cada uno de los diseños usando 4 o 9 bloques.

Aritmética (*V*): se presentan 15 problemas aritméticos en orden de dificultad creciente.

Ensamble de objetos (E): se presentan 4 rompecabezas de cartón al examinado en un formato preestablecido; se pide al examinado que una las piezas para armar algo.

Comprensión (*V*): 18 preguntas que requieren respuestas detalladas se presentan en orden de dificultad creciente.

Símbolos en dígitos (E): 93 casillas que deben llenarse con el símbolo codificado correcto correspondiente al número que aparece sobre la casilla.

Semejanzas (*V*): 14 reactivos del tipo "¿En qué son similares A y B?"

V, subprueba verbal; E, subprueba de ejecución.

mediante la referencia a una tabla especial que viene dentro del manual de administración, la suma de las puntuaciones escaladas de la subpruebas en la Escala verbal puede convertirse en un CI Verbal, la suma de las puntuaciones escaladas de lad subpruebas en la Escala de Desempeño, en un CI de Desempeño, y la suma de las puntuaciones escaladas de las once subpruebas, en una Escala Completa de CI. Éstas son desviaciones CI, expresadas en números en una escala de calificaciones estándar con una media de 100 y una desviación estándar de 15.

Estandarización. La escala WAIS-R se estandarizó en una muestra nacional, cuidadosamente seleccionada, de 1,880 adultos "normales" ubicados en nueve grupos de edad (16 a 17, 18 a 19, 20 a 24, 25 a 34, 35 a 44, 45 a 54, 55 a 64, 65 a 69 y 70 a 74) dentro del rango de 16 a 74 años. La muestra de cada categoría de edad se estratificó por sexo, región geográfica, blanco contra no blanco, educación y ocupación. Se controlaron otras características, tales como residencia urbana contra rural, pero no funcionaron como variables de estratificación. La estandarización de la WAIS-R difirió de la presentada originalmente por la de WAIS en 1955, sobre todo en cuanto a la estratificación de la muestra por grupo étnico y la provisión de muestras más representativas para adultos mayores.

Significado diagnóstico de las calificaciones de Wechsler. Al diseñar la escala WAIS, Wechsler planeó obtener más de un cálculo de la habilidad mental general de una persona. Se consideró que una diferencia significativa entre los CI Verbal y de Ejecución de una persona y el patrón de calificaciones (dispersión) en las once subpruebas era característica de cierto tipo de trastornos mentales y, por lo tanto, potencialmente útil para el diagnóstico clínico. Desafortunadamente, la investigación proporciona poco apoyo para las hipótesis de Wechsler en cuanto a la importancia diagnóstica de la dispersión de la calificación escalada en las diversas subpruebas.

Un problema al tratar de analizar la dispersión de la calificación de las subpruebas en las Escalas de Wechsler es que dichas calificaciones no son muy confiables y algunas subpruebas tienen correlaciones considerables entre sí. En consecuencia, la diferencia entre las calificaciones escaladas de una persona en dos subpruebas dadas debe ser muy grande antes de que pueda considerarse significativa. Las diferencias pronunciadas entre las calificaciones escaladas de subpruebas y entre CI Verbales y de Ejecución tienen cierto valor en el diagnóstico de daño cerebral orgánico y psicopatología y en la diferenciación entre inteligencia y oportunidad. Un CI Verbal considerablemente inferior al CI de Ejecución, por ejemplo, puede ser resultado de una experiencia lingüística limitada o de *carencia cultural*.

WAIS-III. Así como en otras pruebas de habilidades cognoscitivas, el contenido y las normas de los tests de inteligencia de algún modo pierden actualidad con los años. Por ello, una nueva edición de WAIS-R, la WAIS-III, se elaboró a mediados de la década de 1990 y fue publicada por The Psychological Corporation en 1997. Al elaborar la WAIS-III, se prestó particular atención a las subpruebas verbales, tales como Información, Vocabulario y Comprensión, las cuales, debido a que están más sujetas a cambios culturales, se vuelven obsoletas más pronto que otras subpruebas.

Además de las revisiones de las once subpruebas de la WAIS-R, se incluyeron tres nuevas subpruebas en la WAIS-III: Razonamiento de Matriz, Búsqueda de Símbolos y Secuencias de Letras y Números. El Razonamiento de Matriz consiste en una serie de imágenes de cinco formas geométricas; se requiere que los examinados nombren o señalen la forma correcta. En esta subprueba se incluyen cuatro tipos de reactivos, Completar Patrones, Clasificación, Razonamiento por Analogía y Razonamiento Serial. La subprueba de Búsqueda de Símbolos comprende un conjunto de grupos pareados, en el que cada par consiste en un grupo meta y un grupo de búsqueda. Los examinados marcan la casilla adecuada para indicar si cada símbolo meta aparece en el grupo de búsqueda. La subprueba de Secuencias de Letras y Números es una serie de letras y números presentados oralmente en desorden. Los examinados reordenan y repiten la lista diciendo los números en orden ascendente y repitiendo luego las letras en orden alfabético.

El tiempo de aplicación de la WAIS-III es menor que el de la WAIS-R, y el nivel inferior para la mayoría de las subpruebas se ha reducido con el propósito de lograr un mejor cálculo del funcionamiento cognoscitivo de los individuos con retraso mental. Además de los CI tradicionales, Verbal, de Ejecución y de Escala Completa, se obtienen cuatro calificaciones de índice de factores (Comprensión Verbal, Memoria de Trabajo, Organización Perceptual y Velocidad de Procesamiento).

La WAIS-III se estandarizó en una muestra de 2,450 adultos de edades entre 16 y 89 años. La muestra fue estratificada por raza-etnia (blancos, afroamericanos, latinos, otros), sexo, nivel educativo y región geográfica en cada grupo de edad. Las correlaciones entre las calificaciones WAIS-III y las de otras pruebas de la familia Wechsler, así como la cuarta edición de la Escala Stanford-Binet y las Matrices Progresivas de Raven, se incluyen en el manual. También se proporcionan estadísticas basadas en diversos grupos clínicos. Las escalas de la WAIS-III se ajustaron a las normas de la tercera edición de la Escala de Memoria Wechsler (WMS-III), lo que permite efectuar un examen de la relación entre el funcionamiento intelectual de una persona y su memoria.

Escala de Inteligencia para Niños de Wechsler, tercera edición. La Escala de Inteligencia para Niños de Wechsler (WISC), una extensión hacia abajo de la Forma I de la Escala de Wechsler-Bellevue, fue publicada por The Psychological Corporation en 1949. En 1974 se publicó una

revisión de la WISC, la WISC-R, y en 1991 apareció publicada la Escala de Inteligencia para Niños de Wechsler, tercera edición (WISC-III). Esta prueba, diseñada para niños de entre 6 y 16 años 11 meses, consiste en las siguientes seis subpruebas Verbales y siete subpruebas de Ejecución:

SUBPRUEBAS VERBALES	SUBPRUEBAS DE EJECUCION
Información	Completamiento de dibujos e imágenes
Semejanzas	Codificación
Aritmética	Acomodamiento de dibujos
Vocabulario	Diseño de cubos
Comprensión	Ensamble de objetos
Intervalo de dígitos (complementaria)	Búsqueda de símbolos (complementaria)
	Laberintos (complementaria)

Las diez subpruebas principales (no complementarias) pueden administrarse en un lapso de entre 50 y 70 minutos, y las subpruebas complementarias en otros 10 o 15 minutos adicionales. Así como en la WAIS-R, las subpruebas Verbales y de Ejecución de WISC-III se administran alternadamente. Los CI Verbal, de Ejecución y de Escala Completa, basados en la misma escala de calificaciones normalizadas que los de la WAIS-III, se determinan añadiendo la calificación escalada de las cinco subpruebas Verbales y las cinco de Ejecución que se aplican. El WISC-III también puede calificarse para cuatro factores: Comprensión Verbal, Organización Perceptual, Libertad y Distracción y Velocidad de Procesamiento.

La WISC-III fue estandarizada en muestras representativas de estadounidenses de 100 niños y 100 niñas en cada uno de once grupos de edad de los 6 a los 16 años. Las muestras también se estratificaron por región geográfica, nivel educativo de los padres y raza. Se evaluaron otras muestras de niños con WISC-III y con la WAIS-R o la WPPSI-R, dependiendo de sus edades. Las confiabilidades de test-retest de la WISC-III, obtenidas al readministrar la escala después de 4 a 8 semanas, son satisfactorias. Asimismo, se han llevado a cabo diversos estudios de validación con varios grupos clínicos de niños.

Escala de Inteligencia para Nivel Preescolar y Primaria de Wechsler Revisada. Una tercera prueba de Wechsler, la Escala de Inteligencia para Nivel Preescolar y Primaria de Wechsler (WPPSI), fue publicada por The Psychological Corporation en 1967 y una revisión, la WPPSI-R, en 1989. Las seis subpruebas Verbales (V) y las seis de Ejecución (E) de la WPPSI-R, en orden de aplicación, son: Ensamble de Objetos (E), Información (V), Diseño Geométrico (E), Comprensión (V), Diseño de Bloques (E), Aritmética (V), Laberintos (E), Vocabulario (V), Completar Imágenes (E), Semejanzas (V), Piezas con forma de animales (E), y Enunciados (V). Las últimas dos son subpruebas complementarias. Diseñada para niños de entre 3 y 7 años de edad, la WPPSI-R fue estandarizada a fines de la década de 1980 sobre una muestra nacional de niños estadounidenses de entre 3 y 7 años de edad. Estratificar la muestra por género, etnia y nivel educativo y ocupacional de los padres, la hizo más representativa de la población de Estados Unidos en este rango de edad. Del mismo modo que la WAIS-R y la WISC-III, la WPPSI-R produce CI Verbal, de Ejecución y de Escala Completa por separado, con base en una escala de calificación estándar con una media de 100 y desviación estándar de 15.

Escala de Inteligencia de Wechsler Abreviada. En contextos clínicos y educativos, la necesidad de una medición confiable de la inteligencia que pudiera realizarse en forma más rápida que la WAIS-III y la WISC-III condujo a la creación de la Escala de Inteligencia de Wechsler

Abreviada (WASI). Las subpruebas de la WASI se construyeron independientemente de las correspondientes subpruebas de la WAIS-III y la WISC-III, pero en forma paralela. La forma de cuatro subpruebas de la WASI consiste en subpruebas de Vocabulario, Semejanzas, Diseño de Cubos y Razonamiento de Matriz. Las primeras dos constituyen la Escala Verbal y las últimas dos la Escala de Ejecución de la WASI. La forma de dos subpruebas de la WASI incluyen Vocabulario y Razonamiento de Matriz. La forma de cuatro subpruebas requiere de aproximadamente 30 minutos y la de dos subpruebas toma alrededor de 15 minutos en administrarse.

Otras pruebas de inteligencia individuales de rango amplio

Aunque son las pruebas de inteligencia individuales más populares en Estados Unidos, la Stanford-Binet y la de Wechsler de ningún modo son las únicas baterías de amplio rango para evaluar la habilidad mental general. Tampoco son las pruebas más populares de habilidad mental en otros países. De particular relevancia en el Reino Unido son las Escalas de Habilidad Británicas (BAS), que fueron revisadas por The Psychological Corporation y reestandarizadas en Estados Unidos como Escalas de Habilidad Diferencial (DAS).

Escalas de habilidad diferencial. El objetivo de las Escalas de Habilidad Diferencial (DAS) (de The Psychological Corporation) es proporcionar perfiles de habilidad para analizar y diagnosticar problemas de aprendizaje en los niños, evaluar cambios en las habilidades con el tiempo e identificar, seleccionar y clasificar a los niños (de entre $2^{1}/_{2}$ y 17 años de edad) con problemas de aprendizaje. Las DAS consisten en 20 subpruebas, incluyendo 12 subpruebas principales, 5 subpruebas de diagnóstico y 3 subpruebas de aprovechamiento. Las tres subpruebas de aprovechamiento (Habilidades Numéricas, Ortografía, Lectura de Palabras) son útiles para evaluar habilidades académicas básicas, pero las subpruebas centrales y de diagnóstico proporcionan el principal medio de evaluar las cognoscitivas. A cada examinando se le aplican de cuatro a seis subpruebas centrales, de los 2 años 6 meses a los 17 años 11 meses. Se combinan las calificaciones de varias subpruebas centrales para obtener índices generales de Habilidad Verbal, Habilidad de Razonamiento No Verbal y Habilidad Conceptual General, en una escala con una media de 100 y desviación estándar de 15. Aunque las subpruebas de diagnóstico no se usan para calcular los índices de habilidad, proporcionan información útil para comprender las ventajas y deficiencias cognoscitivas del niño.

Las normas de las DAS se basan en 3,475 niños estadounidenses; la muestra fue estratificada por edad, sexo, raza-etnia, educación de los padres, región geográfica e inscripción en educación preescolar. Los niños especiales (con trastornos de aprendizaje, dificultades de habla y lenguaje, retrasados mentales susceptibles de ser educados, superdotados, emocionalmente perturbados, con trastornos sensoriales o motrices) se incluyeron en la muestra.

Prueba Detroit de Habilidad de Aprendizaje. Otra batería relevante es la Prueba Detroit de Habilidad de Aprendizaje (de pro.ed). La administración de la principal edición de esta batería (DTLA-P-2), que fue diseñada para niños de entre 3 y 9 años de edad, dura entre 15 y 20 minutos. Las subpruebas incluyen Articulación, Compaginación Conceptual, Reproducción de Diseños, Secuencia de Dígitos, Dibujar una Persona, Secuencias de Letras, Instrucciones Motoras, Secuencias de Objetos, Instrucciones Orales, Imágenes Fragmentadas, Identificación de Imágenes, Imitación de Enunciados y Relaciones Simbólicas.

La cuarta edición de la Prueba Detroit de Habilidad de Aprendizaje (DTLA-4) fue diseñada para niños de 6 a 17 años y tarda de 50 a 90 minutos en administrarse. Las subpruebas de la DTLA-4 incluyen Palabras Opuestas, Secuencias de Diseño, Imitación de Enunciados, Letras Invertidas, Construcción de Historias, Reproducción de Diseños, Información Básica, Relaciones Simbó-

licas, Secuencias de Palabras y Secuencias de Historias. Calificaciones normalizadas, rangos percentilares y equivalentes de edad pueden determinarse para las diez subpruebas y las pruebas compuestas (General, Nivel Óptimo, Dominio, Teórico). Las pruebas compuestas de Dominio son Verbal, No Verbal, Aumento de la Atención, Reducción de la Atención, Motricidad Aumentada, Motricidad Reducida. Las compuestas teóricas son de inteligencia Fluida y Cristalizada, de Asociación y Cognoscitiva, Simultánea y Sucesiva, Verbal y de Ejecución. La DTLA-4 es un mejoramiento sobre sus predecesoras con respecto a la claridad, facilidad de administración, estandarización, confiabilidad, validez y otras características estadísticas.

Pruebas de inteligencia de Kaufman. La Batería de Kaufman de Evaluación para Niños (K-ABC) (del American Guidance Service) fue diseñada por A. S. Kaufman y N. L. Kaufman con el propósito de evaluar las habilidades de niños de entre $2^1/2$ y $12^1/2$ años de edad para resolver problemas que requieren de un procesamiento mental simultáneo y secuencial. La K-ABC también incluye una Escala de Aprovechamiento para medir habilidades adquiridas en lectura y aritmética. Basada en una extensa investigación sobre neuropsicología y psicología cognoscitiva, la K-ABC fue diseñada especialmente para niños de edad preescolar, menores de edad, y excepcionales. 13 de las 16 subpruebas tipo juego que comprende la K-ABC pueden administrarse en un lapso de 35 a 85 minutos. Las calificaciones se obtienen en cuatro áreas globales: Procesamiento Secuencial, Procesamiento Simultáneo, Compuesta de Procesamiento Mental (Secuencial más Simultánea) y Aprovechamiento.

La muestra de estandarización para la K-ABC, basada en estadísticas registradas en el censo de Estados Unidos de 1980, se estratificó por raza (blanca, negra, latina, asiática, indígena estadounidense) e incluía un grupo representativo de niños excepcionales. Se establecieron normas de rangos percentilares separados por raza y nivel socioeconómico para niños blancos y negros. Los coeficientes de confiabilidad de división por mitades para las cuatro escalas globales en la K-ABC están en los rangos que van del.80 y el.90. En el manual también se da información sobre la validez de constructo, concurrente y predictiva de la prueba.

Otras dos pruebas de inteligencia relevantes diseñadas por A. S. Kaufman y N. L. Kaufman, publicadas por el American Guidance Service, son la Prueba de Inteligencia de Kaufman para Adolescentes y adultos (KAIT) y la Prueba Breve de Inteligencia de Kaufman (K-BIT). Ambas pruebas se basan en la teoría de R. B. Cattell sobre la inteligencia fluida y cristalizada. La KAIT se diseñó para edades entre los 11 y 85+ y toma entre 60 y 90 minutos; la K-BIT está diseñada para edades de 4 a 90 años y dura de 15 a 20 minutos.

Pruebas Woodcock-Johnson III de Habilidades Cognoscitivas. La Woodcock-Johnson III (WJ III) (de Riverside Publishing) consiste en dos baterías co-normalizadas para medir la habilidad intelectual general, habilidades cognoscitivas específicas y el aprovechamiento académico. Una batería, las Pruebas de Habilidades Cognoscitivas Woodcock-Johnson III (WJ III), se basa en la teoría de habilidades cognoscitivas de Cattell-Horn-Carroll (CHC) (vea Woodcock, 1998). Esta batería consiste en una Batería Estándar de diez pruebas y una Batería Ampliada de diez pruebas adicionales. Las pruebas tienen un rango amplio de edad y grado (de 2 a 90+ años; desde jardín de niños hasta la universidad) y una duración de evaluación relativamente breve (aproximadamente cinco minutos por prueba).

Las calificaciones de seis grupos: Verbal-Estándar, de Pensamiento-Estándar, Eficiencia Cognoscitiva-Estándar, Percepción Fonémica, Memoria Funcional y Recuerdo Demorado, se determinan a partir de la Batería Estándar. Las calificaciones de catorce grupos adicionales se obtienen cuando se aplica la Batería Ampliada. Además de las calificaciones en los grupos separa-

dos, se calcula una calificación de Habilidad Intelectual General (GIA) al combinar las calificaciones de las primeras siete pruebas o una calificación GIA (Ampliada) administrando 14 pruebas cognoscitivas. Puede calcularse una calificación de Habilidad Intelectual Breve (BIA) combinando las calificaciones de las pruebas de Comprensión Verbal, Formación de Conceptos y Compaginación Visual. También pueden determinarse calificaciones en los siguientes factores CHC: Comprensión-Conocimiento (Gc), Recuerdo a Largo Plazo (Glr), Pensamiento Visual-Espacial (Gv), Procesamiento Auditivo (Ga), Razonamiento Fluido (Gf), Velocidad de Procesamiento (Gs) y Memoria de Corto Plazo (Gsm).

Sistema de Evaluación Cognoscitiva Das-Naglieri. Otra prueba de inteligencia reciente es el Sistema de Evaluación Cognoscitiva Das-Naglieri (CAS)(Naglieri y Das, 1997)(de Riverside Publishing). El CAS es similar a la Woodcock-Johnson III en cuanto a que está basado en una teoría cognoscitiva y lo publica la misma compañía (Riverside Publishing Co.). Al orientarse hacia niños en edad escolar y adolescentes, el rango de edad del CAS (de 5 años a 17 años 11 meses) es más estrecho que el de la WJ III.

El CAS se diseñó "para proporcionar una medida del procesamiento cognoscitivo que sea justa para niños menores de edad, eficaz para un diagnóstico diferencial y relacionada con la intervención". Se basa en la teoría PASS (Planeación, Atención, Simultánea, Sucesiva) de Das-Naglieri sobre la cognición, descrita en los inicios de ese capítulo, y es adecuada para niños en edad escolar y adolescentes. El tiempo de evaluación es de 40 minutos para la Batería Básica y de 60 minutos para la Batería Estándar. Las subpruebas se agrupan en los cuatro procesos cognoscitivos del modelo PASS:

PLANEACIÓN	SIMULTÁNEA
Números Correspondientes	Matrices No Verbales
Códigos Planeados	Relaciones Verbal-Espaciales
Conexiones Planeadas	Recuerdo de Figuras
ATENCIÓN	**SUCESIVA**
Atención Expresiva	Series de Palabras
Detección de Números	Repetición de Enunciados
Atención Receptiva	Velocidad de Habla (de 5 a 7 años de edad)
	Preguntas de Enunciados (de 8 a 17 años de edad)

La Batería Básica consiste en dos subpruebas, y la Batería Estándar en tres subpruebas, a partir de cada una de estas cuatro categorías.

Además de las calificaciones en las pruebas separadas, las calificaciones normalizadas con una media de 100 y desviación estándar de 15 se obtienen al combinar las calificaciones de todas las escalas. Como la Woodcock-Johnson III, el CAS se estandarizó cuidadosamente y tiene confiabilidades aceptables. Las evidencias de investigación relativas a diversos tipos de validez (de constructo, concurrente, predictiva y discriminante) se registran en el manual de la prueba.

Pruebas no verbales para los discapacitados

Los instrumentos psicométricos que requieren de señalar, manipular objetos o de otra respuesta no verbal, antes que de hablar o escribir, se conocen como *pruebas no verbales*. El desempeño en algunas tareas de estas pruebas puede facilitarse con el lenguaje verbal, pero su uso es mínimo.

El hecho de que las escalas Wechsler contengan medidas verbales y de ejecución separadas las hace más adecuadas que las versiones anteriores de Stanford-Binet para examinar perso-

nas con diferencias físicas, lingüísticas y culturales. Las subpruebas de desempeño Wechsler tienden a ser medidas más precisas de la habilidad mental en niños con problemas de audición y culturalmente distintos, mientras que las subpruebas verbales son medidas más válidas para los ciegos y débiles visuales. Al evaluar a personas ciegas, en ocasiones se ha aplicado una serie de seis pruebas de desempeño especialmente diseñadas conocidas como la Escala de Inteligencia Haptic para Adultos Ciegos, en conjunto con la Escala Verbal del WAIS, como medida de la inteligencia de adultos ciegos y débiles visuales.

Pruebas de una única tarea. Una de las pruebas no verbales más antiguas, el Tablero de Formas Seguin, se introdujo en 1866. Sin embargo, no fue sino hasta la primera parte del siglo XX cuando Knox, Kohs, Porteus y otros psicólogos realizaron serios esfuerzos por estandarizar dichas pruebas. Para medir las habilidades mentales, adicionalmente a muchos tipos de tableros de formas, se han utilizado tareas no verbales como rompecabezas de diversos tipos, el golpeteo secuencial de cubos, problemas de emparejamiento, diseños de cubos, laberintos, dibujo de personas y señalamiento de imágenes.

Los laberintos se han usado en forma extensa tanto en laboratorios psicológicos y clínicas como en varias pruebas estandarizadas. Los Laberintos de Porteus, publicados inicialmente en 1914 y descritos por su diseñador como una medida de la capacidad de previsión y planeación, consisten en un conjunto de laberintos ordenados por dificultad creciente. En cada laberinto se instruye al examinado para que trace el camino más corto entre el punto de partida y el final, sin levantar el lápiz ni entrar en un callejón sin salida. Para quienes padecen algún trastorno verbal, los Laberintos de Porteus son particularmente adecuados como prueba breve (25 minutos), y se han empleado en varias investigaciones y estudios antropológicos sobre los efectos de las drogas y la neurocirugía.

Otra prueba de ejecución no verbal para los discapacitados consiste en diseños de cubos tales como los de las escalas de Wechsler y las Escalas de Habilidad Diferencial. Una de las pruebas más antiguas de este tipo es el Diseño de Cubos de Kohs. Los materiales de la prueba de Kohs son 16 cubos de color y 17 tarjetas con diseños coloreados que el examinado debe copiar. El Diseño de Cubos se consideraba especialmente apropiado para niños con discapacidad de lenguaje y audición, pero ahora su aplicación es muy esporádica.

La Escala de Madurez Mental de Columbia (CMMS) es otra prueba de una única tarea que sólo requiere de señalar. Esta prueba se diseñó originalmente para evaluar niños con parálisis cerebral, pero puede administrarse a otros niños con verbales y motrices disminuidas (discapacidades visuales, trastornos del habla, de la audición, retraso mental) así como a niños hiperactivos. Los materiales de prueba consisten en 92 reactivos (una serie de dibujos) impresos en tarjetas de 15×47.5 cm. Se pide al niño (de entre $3\frac{1}{2}$ y 10 años de edad) que seleccione, a partir de una serie de dibujos presentados en cada tarjeta, la imagen que no pertenece al grupo. Al seleccionar, el niño usa discriminación perceptiva y clasificatoria o habilidades de razonamiento general que incluyen color, forma, tamaño, uso, número, partes faltantes y material simbólico. Los 92 reactivos de la CMMS están dispuestos en ocho niveles traslapados, pero sólo entre 51 y 65 reactivos se aplican de hecho a un examinado determinado. La prueba dura entre 15 y 20 minutos, y las instrucciones se dan en inglés o en español. La ejecución se expresa en términos de calificaciones de desviación de edad desde 50 hasta 150, así como en rangos percentilares, estaninas e índices de madurez.

Baterías de pruebas de ejecución. La primera batería de pruebas de ejecución estandarizadas que se distribuyó comercialmente fue la Escala Pintner-Paterson de Pruebas de Ejecución (1917). Igualmente conocida es la Escala Puntual Arthur de Pruebas de Ejecución, publicada inicialmente por Grace Arthur en 1925. Dos baterías de pruebas de ejecución que se han usado am-

pliamente en niños con discapacidades de habla y de audición, y que todavía están disponibles, son la Escala Leiter de Desempeño Internacional y las Pruebas Hiskey-Nebraska de Aprendizaje. También son interesantes algunas baterías recientemente publicadas, tales como la Prueba Comprensiva de Inteligencia No Verbal, la Prueba de Habilidad No Verbal de Naglieri y la Prueba Universal de Inteligencia No Verbal.

Prueba Hiskey-Nebraska de Habilidad de Aprendizaje. Esta prueba se diseñó específicamente para evaluar las capacidades cognoscitivas de niños con discapacidad auditiva. Consiste en 12 subpruebas no verbales aplicadas mediante instrucciones en pantomima a niños sordos o con instrucciones verbales a niños normales. La prueba se aplica en forma no acelerada y proporciona la edad mental y un cociente de inteligencia. En el momento de escribir el presente libro, la prueba Hiskey-Nebraska se estaba reestructurando en cuanto a sus normas por Slosson Educational Publications, de modo que las características demográficas de la muestra de estandarización se ajustaran a las de la población actual de Estados Unidos.

Leiter-R. La versión revisada de la Escala Leiter de Desempeño Internacional (Leiter-R) (de Stoelting) se promueve como una medida de las habilidades cognoscitivas que es justa para la cultura y adecuada para personas de varios contextos culturales. Tiene un rango de edad de 2 a 21 años y puede administrarse sin lenguaje verbal a niños con problemas de audición o con otros trastornos lingüísticos de expresión o de recepción y culturalmente diferentes, con discapacidades motrices, autistas e incluso a niños superdotados. Se solicita a los examinados que unan una serie de cartas de respuesta coloreadas con las ilustraciones correspondientes presentadas en un caballete. Las cuatro subpruebas de Razonamiento y las seis de Visualización de la batería de Visualización y Razonamiento requieren de un total de 40 minutos para administrar. Las ocho subpruebas de Memoria y las dos de Atención en la correspondiente batería toman 35 minutos. Evaluar el CI o LD/ADHD puede realizarse en 25 minutos administrando una batería incompleta; la Evaluación de Selección de Superdotados requiere de 35 minutos. La Leiter-R fue estandarizada en 1993 en 1,719 niños típicos y 692 atípicos de entre 2 y 12 años de edad. Las evidencias de confiabilidad y validez registradas en el manual indican que Leiter-R es un instrumento bastante seguro psicométricamente.

Prueba Comprensiva de Inteligencia No Verbal. La Prueba Comprensiva de Inteligencia No Verbal (CTONI)(de pro.ed) tiene un rango de edad muy amplio (de 6-0 a 90-11) y puede administrarse en alrededor de una hora. Es particularmente apropiada para calcular la inteligencia de niños y adultos con problemas de lenguaje o de habilidades motoras finas. Pueden ser personas que hablen una lengua distinta al inglés, tengan desventajas socioeconómicas o sean sordos, o que sufran algún trastorno del lenguaje, una discapacidad motora o un problema neurológico.

Las seis subpruebas de CTONI son Analogías Pictóricas, Categorías de Imágenes, Secuencias Pictóricas, Analogías Geométricas, Categorías Geométricas y Secuencias Geométricas. Estas subpruebas se diseñaron para medir el razonamiento analógico, las calificaciones categóricas y las habilidades de razonamiento secuencial, como lo revelan las respuestas a las imágenes de objetos familiares (animales, personas, juguetes y reactivos similares) y diseños geométricos (dibujos, esbozos inusuales, etc.). Así como en la prueba Leiter-R, en la CTONI los examinados indican sus respuestas señalando opciones alternativas. Las respuestas se califican entonces y las calificaciones se combinan para obtener tres cocientes compuestos: Cociente de Inteligencia No Verbal (CINV), Cociente de Inteligencia No Verbal Pictórica (CINVP)y Cociente de Inteligencia No Verbal Geométrica (CINVG).

La CTONI se estandarizó en 25 estados de Estados Unidos, Canadá y Panamá. Aunque bastante pequeñas, las muestras se estratificaron por género, regiones geográficas, etnia, raza, residen-

cia urbana-rural y discapacidad. Esta prueba reporta coeficientes de confiabilidad de .80 o mayores, y en el manual también se da cierta evidencia para la validez con referencia a criterios, de contenido y de constructo. De especial relevancia son los esfuerzos de los diseñadores de la CTONI para detectar y eliminar sesgos culturales, de género, raciales y lingüísticos en los reactivos.

Prueba Universal de Inteligencia No Verbal (UNIT). A diferencia de muchas pruebas no verbales que se caracterizan por una modalidad no verbal ya sea en la administración o bien en los formatos de respuesta, a fin de garantizar la justicia sin importar la cultura, la etnia, el género o la habilidad auditiva, la UNIT se desarrolló con ambas modalidades simultáneamente. La administración de esta batería de prueba implica múltiples modos de respuesta, incluyendo el uso de herramientas de manipulación, lápiz y papel así como señalamiento. El examinador usa ocho gestos universales de manos y cuerpo para explicar las tareas de la prueba al examinado. Además de estos gestos, la aplicación de la prueba incluye demostraciones por parte del examinador, reactivos de muestra, respuestas correctoras, reactivos de transición en puntos de verificación y reactivos que no permiten la retroalimentación del examinador.

La UNIT (de Riverside Publishing) es adecuada para individuos con impedimentos de habla, lenguaje o audición, así como para aquellos que no son comunicativos verbalmente o provienen de distintos contextos culturales o lingüísticos. Los materiales de la prueba se seleccionaron considerando que fueran relativamente independientes de las culturas particulares y de interés para niños con distintos ambientes culturales.

Hay seis subpruebas en la Batería Ampliada de la UNIT: Memoria Simbólica, Memoria de Objeto, Memoria Espacial, Razonamiento Analógico, Diseño de Cubos y Laberintos. Las puntuaciones crudas de estas subpruebas se convierten en calificaciones escaladas con una media de 10 y desviación estándar de 3. También se definen cinco cocientes, Cociente de Inteligencia de Escala Completa (FSIQ), Cociente de Memoria (MQ), Cociente de Razonamiento (RQ), Cociente Simbólico (SQ) y Cociente No Simbólico (NSQ), a partir de la combinación de las calificaciones obtenidas en seis subpruebas de la Batería Ampliada o en cuatro subpruebas de la Batería Estándar. La Batería Ampliada toma 45 minutos en su administración, mientras que la Batería Estándar sólo 30 minutos. Una Batería Abreviada de dos subpruebas, que puede usarse para seleccionar, tarda de 10 a 15 minutos en completarse.

La UNIT se estandarizó hacia mediados de la década de 1990 en una muestra nacional de 2100 niños y adolescentes (de entre 5 años y 17 años 11 meses de edad). Los datos de confiabilidad son satisfactorios, y la evidencia de investigación que corresponde a la validez concurrente, predictiva y discriminativa de este instrumento se proporciona en el manual.

TESTS DE INTELIGENCIA COLECTIVOS

Durante la segunda década del siglo XX, Lewis Terman impartía habitualmente un curso en la Universidad de Stanford sobre la Escala de Inteligencia de Stanford-Binet. Según se informó, en una sección de este curso un estudiante, Arthur Otis, tuvo la idea de adaptar tareas seleccionadas de la Stanford-Binet a un formato de lápiz y papel. Poco después, muchas de las tareas adaptadas por Otis y otros autores se combinaron como la primera prueba de inteligencia colectiva, el Examen Alfa del Ejército.

Los exámenes Alfa y Beta del Ejército, una prueba no verbal lingüística para no angloparlantes y analfabetos, se administraron a casi dos millones de reclutas del ejército estadounidense durante y después de la Primera Guerra Mundial con propósitos de selección militar y clasificación de puestos. El Examen Alfa del ejército consistía en reactivos que incluían analogías, problemas aritméticos, completamiento de series de números, sinónimos y antónimos, análisis de cubos, símbolos en dígitos, información y juicio práctico. Esto propició la aparición de

otras pruebas colectivas de inteligencia y de aptitudes académicas, y funcionó como su modelo después de la guerra. Arthur Otis y otros psicólogos empezaron a publicar sus propias pruebas de inteligencia colectivas, y hacia la década de 1930 había disponibles comercialmente muchos más instrumentos de este tipo.

Una prueba colectiva grupal de inteligencia típica puede constar de un conjunto de preguntas de opción múltiple dispuestas en un formato en espiral-ómnibus o de una serie de subpruebas en momentos separados. En el *formato colectivo en espiral* se mezclan los diversos tipos de reactivos que comprende la prueba y se ordenan por dificultad creciente; los reactivos con el mismo grado de dificultad se agrupan juntos.

Aplicación, calificación e informes

Las pruebas de inteligencia colectivas pueden administrar a pequeñas cantidades de niños desde los 5 o 6 años de edad o a grupos más numerosos de adultos. Al evaluar niños pequeños, los examinadores deben tener particular cuidado en asegurarse de que los examinados comprendan las instrucciones, pasen a la página correcta, comiencen y terminen a tiempo, entre otros aspectos. Al calificar pruebas de inteligencia colectivas, las puntuaciones crudas, ya sea parciales o globales, pueden convertirse en rangos percentilares, calificaciones estándar y otras unidades numéricas.

Incluso más que en pruebas individuales, las calificaciones de pruebas aplicadas de manera colectiva deben interpretarse con precaución, tomando en cuenta otros datos (grados escolares e información obtenida mediante entrevistas o la observación) sobre el examinado. El informe 7.1 que aparece en la página siguiente ilustra la manera en que los hallazgos de una prueba colectiva grupal de inteligencia pueden registrarse e interpretarse, junto con más información relevante sobre el examinado. Asimismo, pueden prepararse perfiles de calificaciones interpretativos a través de un servicio de calificación de pruebas. Los examinados con calificaciones muy bajas deben ser sometidos a otras pruebas, de preferencia individuales, antes de tomar decisiones sobre diagnóstico o colocación.

Ejemplos de pruebas de inteligencia colectivas

Tres de las pruebas de inteligencia colectivas más populares son la Prueba Otis-Lennon de Habilidad Escolar, la Prueba de Aptitudes Cognoscitivas y la Prueba de Personal Wonderlic.

Prueba Otis-Lennon de Habilidad Escolar. Esta prueba (de Harcourt Brace) es una revisión de las Pruebas Autoadministrables de Otis de Habilidad Mental (OLSAT), la Prueba Otis-Lennon de Habilidad Mental y las Pruebas Otis de Habilidad Mental de Calificación Rápida. Igual que sus predecesoras, la séptima edición de la OLSAT consiste en diversos reactivos de imágenes, verbales, de figuras y cuantitativos, a fin de medir Comprensión Verbal, Razonamiento Verbal, Razonamiento de Imágenes, Razonamiento de Figuras y Razonamiento Cuantitativo, desde la etapa preescolar hasta el 12° grado. Hay dos formas y siete niveles de la OLSAT, cada una de las cuales puede administrarse en 60 o 75 minutos. Las normas, que se basan en una muestra nacional amplia, se expresan como rangos percentilares, calificaciones estaninas y NCE por grado. De igual manera pueden realizarse comparaciones entre la habilidad y el logro cuando se aplica la OLSAT con la Serie de Pruebas de Aprovechamiento de Stanford, en su novena edición.

Prueba de Habilidades Cognoscitivas. La quinta edición de la Prueba de Habilidades Cognoscitivas (CogAT) (de Riverside Publishing) evalúa las habilidades de los niños para razonar y resolver problemas usando símbolos verbales, cuantitativos y espaciales (no verbales). La CogAT es una prueba de niveles múltiples, con los niveles 1 y 2 para los grados K-3 y niveles de la A a la H para los grados 3-12; su administración dura aproximadamente 90 minutos. Cada ni-

REPORTE 7.1 Resultados de una prueba colectiva de inteligencia

Nombre del examinado: Jane N. Brown
Fecha de nacimiento: 11 de marzo de 1980
Dirección: 12449 Mount Olive Street
Thousand Oaks, CA

Sexo: Femenino
Edad: 21 años, 11 meses
Escolaridad: Licenciatura universitaria
Fecha de aplicación: 15 de abril de 2002

Prueba aplicada: Prueba Otis-Lennon de Habilidad Escolar, Forma avanzada R

Jane Brown, una joven de altura y peso promedios (1.65 m, 60 kg), se ofreció como voluntaria para someterse a la prueba de inteligencia debido a un interés personal en sus habilidades mentales y como un favor hacia el examinador. La prueba se administró como una tarea en Psicología 405 (Evaluación Psicológica) en Western College durante el semestre de primavera de 2002.

En la época del examen, Jane estaba en su último semestre de la especialización de contabilidad. Informó que su promedio de graduación era de 3.2, y señaló que le gustaría ir a la escuela de posgrado en administración para obtener un grado de maestría, pero que inmediatamente después de su graduación planeaba trabajar de tiempo completo en una empresa de contabilidad en el área de Los Ángeles.

El padre de Jane tiene un grado universitario, y su madre terminó dos años de educación universitaria. Ambos trabajan en el negocio familiar, una empresa de asesoría fiscal. Jane informa haber obtenido buenas calificaciones (B y A) en toda su educación escolar, pero confesó que "¡No soy ninguna académica!" Parece ser muy práctica en cuanto a sus intereses, como lo indica no sólo la licenciatura que eligió, sino también sus planes y otros comentarios que hizo al examinador.

Jane mostró un interés moderado en las preguntas de la prueba y se mostró relajada pero involucrada durante el proceso de evaluación. Trabajó con atención y sin interrupciones durante todos los 40 minutos. Las condiciones de la prueba fueron buenas; no hubo interrupciones ni distracciones.

Resultados e interpretación de la prueba

Jane terminó todas las preguntas de la prueba durante el tiempo estipulado (40 minutos). Obtuvo las siguientes calificaciones en la prueba Otis-Lennon:

Puntuación cruda = 65
Índice de habilidad escolar = 116
Rango percentilar (grupo de 18+ años de edad) = 84
Estanina = (grupo de 18+ años de edad) = 7

Estas calificaciones corresponden aproximadamente al promedio para los estudiantes que han completado la licenciatura universitaria, e indican una habilidad intelectual general en el rango del "Promedio superior" para la población general. Un breve análisis de los 16 reactivos que contestó Jane en forma incorrecta indica que en cierta medida tiene mayor dificultad con el razonamiento no verbal que con el verbal. Sin embargo, no hubo ningún patrón significativo en los errores que cometió; en general fueron bastante aleatorios.

Conclusiones y recomendaciones

En una entrevista posterior al examen, Jane señaló que había realizado su mejor esfuerzo en la prueba y que no tenía prisa por terminar a tiempo. Acabó la prueba en 35 minutos y dedicó los restantes 5 a verificar sus respuestas. Afirmó que el Índice de habilidad escolar, que el examinador le comunicó, se encontraba dentro del rango de 5 puntos de una calificación de CI que obtuvo en una prueba de inteligencia que había realizado en la preparatoria. No pudo recordar el nombre de dicha prueba.

Tomando en cuenta las condiciones de la evaluación, el comportamiento que se observó en el examinando y sus afirmaciones después de la prueba, los resultados se consideran válidos en este momento. Los planes y aspiraciones profesionales de Jane parecen adecuados a su habilidad intelectual, aunque tal vez tendrá que esforzarse con diligencia para obtener una maestría en alguna institución de prestigio.

Laura F. Green

Laura F. Green
Pasante de la Licenciatura en Psicología
Western College

vel contiene una Batería Verbal, una Batería Cuantitativa y una Batería No Verbal que consisten en dos o tres subpruebas. Las calificaciones separadas obtenidas en las tres baterías y una calificación compuesta general pueden convertirse a diversos tipos de calificación normalizadas (calificaciones de edad estándar, rangos percentilares de grado y edad nacionales, calificaciones estaninas de grado y edad, y equivalentes de curva normal) con base en una estandarización nacional llevada a cabo en 1992.

Prueba de Personal Wonderlic. La Prueba de Personal Wonderlic (de Wonderlic) es un instrumento breve (12 minutos) de 50 reactivos basada originalmente en la Prueba Autoaplicable de Otis de Habilidad Mental. Las preguntas de la Wonderlic, cuyos ejemplos se muestran en la figura 7.3, consisten en analogías, definiciones, problemas lógicos y aritméticos, relaciones espaciales, comparaciones entre palabras y ubicación de dirección. Esta prueba se ha usado ampliamente como herramienta de selección en situaciones laborales durante muchos años, y la

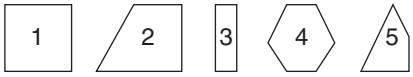
Observe la lista de números que se presenta a continuación. ¿Qué número debe seguir?

8 4 2 1 1/2 1/4 ?

Suponiendo que las dos primeras afirmaciones son ciertas, ¿la última de ellas es:
(1) cierta, (2) falsa, (3) incierta?

El niño juega béisbol. Todos los jugadores de béisbol usan sombrero.
El niño usa sombrero.

Una de las siguientes figuras numeradas es la que más se diferencia de las otras. ¿Qué número tiene dicha figura?



Un tren recorre 60 metros en 1/5 de segundo. A la misma velocidad, ¿qué distancia recorrerá en tres segundos?
¿Cuántos de los seis pares de cifras de la siguiente lista son duplicados exactos?

3421	1243
21212	21212
558956	558956
10120210	10120210
612986896	612986896
356471201	356571201

Las horas de luz diurna y de oscuridad en SEPTIEMBRE son más similares a las horas de luz diurna y oscuridad de:

(1) Junio (2) Marzo (3) Mayo (4) Noviembre

FIGURA 7.3 Muestra de reactivos de la Prueba de Personal Wonderlic.
(Reproducida con autorización de Wonderlic Personnel Test, Inc., Libertyville, IL.)

investigación indica que es un dispositivo justo y válido para la selección en un amplio rango de puestos. A pesar de la brevedad de la prueba Wonderlic, hay registros de que sus coeficientes de confiabilidad y sus correlaciones con calificaciones de otras medidas de inteligencia llegan a .90.

Pruebas de inteligencia colectivas grupales no verbales y justas para las culturas

Las pruebas de ejecución aplicables individualmente y diseñadas como medidas de las habilidades intelectuales de personas con desventajas lingüísticas o culturales ya se trataron en este capítulo. Ahora veamos cómo también se han elaborado instrumentos complementarios que pueden administrarse en forma colectiva para evaluar la inteligencia de individuos con desventajas físicas o culturales. El antecesor de estas pruebas no verbales fue el Examen Army Beta del Ejército aplicado a los reclutas estadounidenses de la Primera Guerra Mundial, el cual incluía tareas como análisis de cubos, símbolos en dígitos, construcciones geométricas, laberintos y completamiento de imágenes. Esta prueba también resultó útil para evaluar a trabajadores civiles no capacitados y fue actualizada, reestandarizada y reeditada, por The Psychological Corporation, en 1978 como Examen Beta Revisado, segunda edición, y de nuevo en 1999 como Beta III.

Test de Dibujo de Goodenough-Harris. Otra prueba no verbal adecuada para su administración colectiva (o individual) es el Test de Dibujo Goodenough-Harris (de The Psychological Corporation). A diferencia de Beta III, que es una prueba de tareas múltiples, la Goodenough-Harris sólo requiere que el examinando realice dibujos de un hombre, una mujer y de sí mismo. Más que calificarse por su mérito artístico, los dibujos se evalúan comparándolos con los doce dibujos modelo y por la presencia de 73 características específicas (por ejemplo, detalles corporales y de vestimenta, proporcionalidad de cabeza y tronco). La prueba no tiene límite de tiempo, pero suele durar entre 10 y 15 minutos. Las normas para niños de entre 3 y 15 años de edad se registran como calificaciones y rangos percentilares, en forma separada por sexo. También es interesante un sistema de calificación cuantitativa, el Dibuja una Persona: QSS, desarrollado por J. A. Naglieri que hace más objetiva la calificación del dibujo de personas.

Pruebas justas para las culturas. Durante muchos años, los diseñadores de pruebas de inteligencia han sido atacados por la crítica de que estos instrumentos están repletos de sesgos culturales de la sociedad occidental de clase media. Goodenough y Harris tenían la esperanza de que su prueba mediría la inteligencia básica relativamente al margen de influencias culturales, pero ha quedado claro que la tarea de dibujar una figura humana está considerablemente alterada por las experiencias socioculturales específicas. Ha habido varios intentos relevantes por elaborar una prueba de inteligencia independiente de la cultura, pero esos esfuerzos no han tenido éxito alguno. Por consiguiente, el objetivo se modificó después por el de desarrollar una prueba de inteligencia que resultara justa para las culturas. Al diseñar una *prueba de inteligencia justa para las culturas*, se intenta usar sólo reactivos relacionados con experiencias comunes a un amplio espectro de culturas. Se omiten reactivos que incluyan construcciones lingüísticas particulares y otras tareas embebidas de cultura, tales como la velocidad de respuesta. En este sentido, el test de Goodenough-Harris es culturalmente justo. Otras pruebas muy utilizadas que probablemente también están cerca de resultar justas para la cultura son las Matrices Progresivas de Raven y la prueba de Inteligencia Justa para la Cultura.

Matrices Progresivas de Raven. Esta prueba, que puede administrarse ya sea en forma individual o colectiva, demanda al examinado indicar cuál de diversas figuras o diseños pertenece a

una matriz dada. Desarrollada en Gran Bretaña como una medida del factor de inteligencia general de Spearman, la prueba de Raven está disponible en The Psychological Corporation en formas de matrices progresivas Estándar, Coloreada y Avanzada. La Forma Estándar, para edades de 6 a 80 años, incluye cinco conjuntos en blanco y negro de 12 problemas cada uno y se termina en un lapso de 20 a 45 minutos. La Forma Coloreada, para niños de 5 a 11 años, individuos de edad avanzada y personas mental y físicamente impedidas, toma de 15 a 30 minutos en terminarse. La Forma Avanzada tiene un rango de los 11 años a la edad adulta y dura entre 40 y 60 minutos. Las normas más recientes, basadas en muestras británicas y estadounidenses, están disponibles en la Forma Avanzada, pero las tres formas requieren de reestandarización.

Una prueba similar, pero más reciente que las Matrices Progresivas de Raven, es la Prueba de Analogías de Matriz-Forma Ampliada. Consiste en reactivos de razonamiento no verbal en cuatro categorías: Completamiento de Patrones, Razonamiento por Analogía, Razonamiento en Serie y Visualización Espacial. Los examinados (de entre 5 y 17 años de edad) son evaluados en un lapso de 20 a 25 minutos con 64 diseños abstractos del tipo de la matriz progresiva estándar, con un diseño por página. Las normas están basadas en una muestra representativa numerosa de individuos de 5 a 17 años de edad, residentes en Estados Unidos. Las puntuaciones crudas se convierten en calificaciones estándar, rangos percentilares y estaninas por intervalos de medio año y en equivalentes de edad de los 5 años a los 17 años 11 meses. The Psychological Corporation también tiene disponible una Prueba de Analogías de Matriz-Forma Abreviada, que consiste en 34 reactivos.

Pruebas de Inteligencia Justas para las Culturas. Estas pruebas (de IPAT) están compuestas por tres escalas: la Escala 1, para niños de 4 a 8 años de edad y adultos retrasados mentales; la Escala 2, para niños entre 8 y 14 años y adultos de inteligencia promedio, y la Escala 3 para estudiantes universitarios, ejecutivos y otras personas de inteligencia superior al promedio. Cada escala contiene cuatro subpruebas (Series, Clasificaciones, Matrices y Condiciones) para medir la habilidad para percibir relaciones. Además de estas cuatro subpruebas justas para las culturas, la Escala 1 contiene cuatro subpruebas para evaluar información cultural y comprensión verbal. La Escala 1 no tiene límite de tiempo, pero toma alrededor de 22 minutos resolverla; las escalas 2 y 3 se llevan $12^{1}/_{2}$ minutos cada una.

Prueba Naglieri de Habilidad No Verbal. La Prueba Naglieri de Habilidad No Verbal-Forma Multinivel (NNAT) (The Psychological Corporation) es similar a la de Raven en cuanto a su diseño de matrices. En la figura 7.4 se presentan ejemplos de los reactivos de la NNAT. El objetivo de esta prueba, así como el de otras pruebas no verbales, es proporcionar una medida no sesgada de la habilidad mental general de individuos con habilidades limitadas para la lengua inglesa o con otros problemas de aprendizaje. La NNAT-Forma Multinivel es adecuada para alumnos desde preescolar hasta el 12° grado, y tarda aproximadamente 30 minutos en completarse. Se proporcionan las Calificaciones del Índice de Habilidad No Verbal y otras calificaciones establecidas basadas en una muestra de estandarización de más de cien mil alumnos.

Instrumentos tales como las Matrices Progresivas de Raven, las Pruebas de Inteligencia Justas para las Culturas, la Prueba Naglieri de Habilidad No Verbal y la Prueba de Inteligencia No Verbal Universal, representan esfuerzos encomiables por elaborar pruebas en que los distintos grupos culturales obtengan calificaciones iguales. No obstante, ahora se reconoce que tal vez sea imposible construir una prueba que mida las habilidades cognoscitivas independientemente de la experiencia. En cualquier caso, los resultados de la investigación realizada en países en vías de desarrollo indica que las diferencias en cuanto a los índices generales de alfabetismo y escolaridad son más importantes que la lengua, el país, la raza o la etnia para determinar diferencias "culturales" en las calificaciones de pruebas de inteligencia (vea Frisby, 1999).

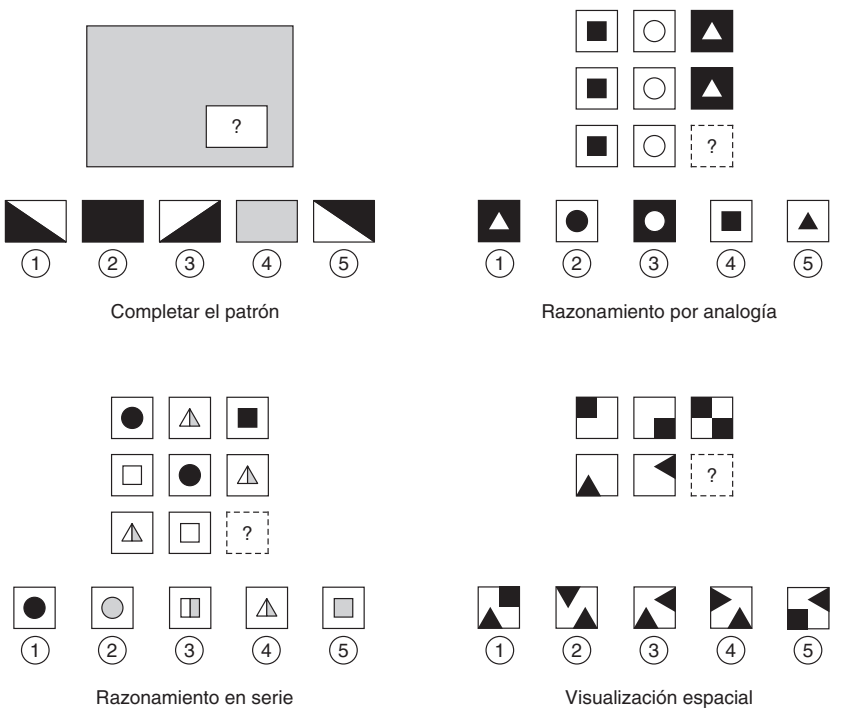


FIGURA 7.4 Ejemplos de reactivos de la Prueba Naglieri de Habilidad No Verbal-Forma Multinivel.

(Derechos Reservados © 1996 por Harcourt, Inc. Reproducido con autorización. Reservados todos los derechos, "Naglieri Nonverbal Ability Test" y "NNAT" son marcas registradas propiedad de The Psychological Corporation e inscritas en Estados Unidos de Norteamérica y otras jurisdicciones.)

Como se observa en la Prueba de Inteligencia No Verbal Universal, continúan los esfuerzos por desarrollar pruebas de habilidades cognoscitivas que sean justas para personas de distintas culturas, pero ello no significa que muchas antiguas pruebas de inteligencia muy sesgadas culturalmente deban abandonarse. Es digno de mención que en otros países el mercado de pruebas de inteligencia tradicionales es mucho mayor que el de pruebas justas para las culturas (Oakland y Hu, 1993). ¡Aparentemente, las personas de países no occidentales se preocupan menos que los estadounidenses de clase media por lo justo para las culturas que resulten las pruebas tradicionales de inteligencia tipo Binet!

Pruebas de aptitud académica y de admisión

Muchas pruebas de inteligencia colectivas se han diseñado específicamente con el propósito de medir la aptitud para el trabajo académico y se conocen como *pruebas de aptitud académica.* Algunas pruebas de inteligencia en grupo tienen un enfoque más amplio que éste, pero aun así su contenido es similar al de las medidas de habilidad académica: tienen un gran contenido de reactivos verbales, numéricos y otros de tipo escolar.

A lo largo del tiempo se han usado muchas pruebas distintas con propósitos de admisión a las universidades, incluyendo el Examen Psicológico del Consejo Estadounidense sobre Educación

(ACE), las Pruebas de Habilidad Universitaria y Escolar (SCAT), la Prueba de Habilidad Académica del Consejo de Exámenes de Ingreso a la Universidad (ahora denominada Prueba de Evaluación Académica, o SAT), y la Evaluación del Programa de Pruebas Universitarias Estadounidenses (ACT). Debido a su extenso uso, se describirán con cierto detalle las últimas dos de estas baterías.

Prueba de Evaluación Académica (SAT). Antes de 1994, la SAT, anteriormente denominada Prueba de Habilidad Académica, consistía en dos secciones que producían dos calificaciones: Verbal (SAT-V) y Matemática (SAT-M). La sección verbal estaba compuesta por reactivos de analogías verbales, antónimos, información, comprensión de lectura y completamiento de enunciados; la sección matemática consistía en reactivos de aritmética, álgebra, geometría, cuadros y gráficas y razonamiento lógico. Ambas secciones se calificaron en una escala estándar con una media de 500 y desviación estándar de 100, con las calificaciones en un rango de 200 a 800. Aunque cada año se desarrollaron versiones nuevas de la SAT, las calificaciones de cada nueva forma se escalaron hacia el grupo de estandarización de 1941. Este grupo estuvo constituido por diez mil alumnos del noreste de Estados Unidos, en su mayoría varones de raza blanca y con nivel de ingresos alto, los cuales estaban solicitando su admisión a las escuelas de la Ivy League. Como es comprensible, los estudiantes de preparatoria de principios de la década de 1990 obtuvieron calificaciones algo inferiores a la media de 500 lograda por este grupo.

La versión actual de la SAT, que se administró primero a nivel nacional en marzo de 1994, está compuesta por dos partes, SAT I: Razonamiento, y SAT II: Pruebas de Materia. SAT I consiste en secciones de Razonamiento Verbal y Razonamiento Matemático con una duración de 75 minutos cada una. La sección de Razonamiento consta de 78 reactivos de opciones múltiples en Analogías, Completamiento de Enunciados y Lectura Crítica. La sección de Razonamiento Matemático está formada por 60 reactivos en Matemáticas Regulares, Comparaciones Cuantitativas y Respuestas Producidas por el Alumno. Se pide a los examinados que lleven al examen su propia calculadora de bolsillo, de modo que puedan calcular las respuestas de las subpruebas matemáticas.

Así como en versiones anteriores de la SAT, las puntuaciones crudas de las Pruebas de Razonamiento se convierten a una escala de calificación estándar que tiene una media de 500 y una desviación estándar de 100. La calificación de la versión revisada de la SAT, la Prueba de Evaluación Académica, se basa en el desempeño de más de un millón de estudiantes que se sometieron a la prueba en 1994. Las calificaciones se *recalcularon* para reflejar la población estudiantil mayor y más diversa de la actualidad, lo que dio como resultado un aumento de la calificación promedio de Razonamiento Verbal de aproximadamente 80 puntos, y de la calificación promedio del Razonamiento Matemático en alrededor de 20 puntos. Además de las calificaciones estándar en las Pruebas de Razonamiento Verbal y Matemático, un informe de calificaciones de SAT da puntuaciones crudas y rangos percentilares para cada subprueba, rangos de calificaciones basadas en error estándar de medida de las pruebas y equivalentes de percentiles nacionales y estatales para estudiantes universitarios del último año. Los resultados de múltiples estudios indican que la SAT-I es un predictor válido para el desempeño en la universidad, específicamente de los promedios de grado durante el primer semestre universitario, pero también predice con eficacia los promedios posteriores y el desempeño en otros exámenes académicos.

Las 20 Pruebas de Materia SAT pertenecen a cinco áreas generales: Inglés, Historia y Estudios Sociales, Matemáticas, Ciencias y Lenguas. Se obtiene una muestra directa de las de redacción del examinado, y también se administran preguntas de opción múltiple sobre inglés escrito, dicción y expresión lógica. Así como sucede con las calificaciones de la SAT-I, las calificaciones de la SAT-II se registran en una escala de calificación estándar con una media de 500 y desviación estándar de 100.

Pruebas Universitarias Estadounidenses. El segundo examen de admisión a la universidad más usado es el constituido por las Pruebas Universitarias Estadounidenses (ACT), el cual se aplica cinco veces al año tanto en Estados Unidos como en otros países. Hay cuatro subpruebas en las ACT: Inglés, Matemáticas, Lectura y Razonamiento Científico. A quienes se someten a este examen se les entregan calificaciones de las cuatro subpruebas, así como una calificación compuesta (el promedio de las cuatro subpruebas redondeado al entero más cercano) y siete subcalificaciones. Las calificaciones compuestas y las de las subpruebas van de 1 a 36, con una media de 18; las siete subcalificaciones están entre 1 y 18, con una media de 9. Las confiabilidades de las cuatro subpruebas van desde .78 para Razonamiento Científico hasta .91 para Inglés, con coeficientes de consistencia interna algo más elevados que los coeficientes de formas paralelas. Como podría esperarse debido a su extensión más corta, las confiabilidades de las subcalificaciones son inferiores a las de las subpruebas, y están entre .67 para Geometría Plana/Trigonometría y .85 para Uso/Mecánica del Inglés.

Exámenes del Registro de Graduados. La prueba más popular para admisión en una escuela de posgrado es el Examen del Registro de Graduados (GRE). Consiste en una Prueba General que mide la aptitud para el trabajo de posgrado y una serie de Pruebas de Materia que miden el aprovechamiento en una materia en particular. La Prueba General contiene tres secciones calificadas: una sección Verbal (V) de 30 minutos con 30 preguntas, una sección Cuantitativa de 45 minutos con 28 preguntas, y una sección Analítica (A) de 60 minutos con 35 preguntas. Los reactivos de la sección Verbal consisten en analogías, antónimos, completamiento de enunciados y comprensión de lectura. Los reactivos de la sección Cuantitativa incluyen comparación cuantitativa, cuantitativa discreta y problemas de interpretación de datos. Los reactivos de las pruebas analíticas constan de ejercicios de razonamiento analítico y de razonamiento lógico. La Prueba General produce calificaciones separadas: Verbal (GRE-V), Cuantitativa (GRE-Q), y Análisis (GRE-A), con la misma escala de calificación estándar que la SAT. Las Pruebas de Materia GRE son exámenes de tres horas sobre materias en particular (vea el capítulo 6).

RESUMEN

Las definiciones y teorías de *inteligencia*, un antiguo término latino reintroducido durante el siglo XIX, se dieron en abundancia en el siglo pasado. Entre las teorías más prominentes sobre la inteligencia figuran las relacionadas con el enfoque psicométrico (Spearman, Thurstone, Guilford, Vernon, Cattell), el enfoque del desarrollo (Piaget), y el enfoque del procesamiento de información (Sternberg, Gardner, Das y Naglieri).

Varias pruebas sensoriomotrices se usaron en los primeros intentos por evaluar la inteligencia, pero la primera prueba práctica de inteligencia fue elaborada por Alfred Binet y Théodore Simon durante la primera década del siglo XX. La Escala Binet-Simon, una serie de tareas relacionadas con la escuela y ordenadas por dificultad creciente, proporciona una calificación de edad mental para cada examinando. Entre las múltiples traducciones y revisiones de la Escala Binet-Simon, la más popular fue la Escala de Inteligencia Stanford-Binet, que fue publicada por primera vez en 1916 y revisada en 1937 y 1960, y cuyo autor fue Lewis Terman. La prueba producía un valor llamado razón de CI, definida como

$$\text{CI} = 100\left(\frac{\text{MA}}{\text{CA}}\right)$$

aunque una desviación del CI podía también calcularse en la revisión de 1960.

La cuarta edición de la Escala Stanford-Binet representó un considerable alejamiento de las ediciones previas. La elaboración de la cuarta edición, que incluía una teoría y una metodología psicométrica más complejas, proporcionaba calificaciones separadas en 15 pruebas y cuatro áreas, así como una calificación compuesta. El énfasis al diseñar la cuarta edición residía no sólo en identificar el retraso mental, sino también en proporcionar información para diagnosticar causas específicas de problemas de aprendizaje.

Durante muchos años, las primeras ediciones de la Escala Stanford-Binet funcionaron como un patrón contra el cual se comparaban otras pruebas de inteligencia. Sin embargo, en la segunda mitad del siglo pasado, las escalas de inteligencia de Wechsler se volvieron más populares. A diferencia de las subpruebas de la Stanford-Binet, que se agrupan de acuerdo con niveles de edad, las subpruebas de las escalas Wechsler se dividen en aproximadamente diez categorías de acuerdo con su contenido. Asimismo, las calificaciones de las escalas Wechsler proporcionan tres tipos de CI de desviación: Verbal, de Ejecución y de Escala Completa. Sumado a los tres CI, el patrón de calificaciones escaladas de las subpruebas en las pruebas de Wechsler puede proporcionar información clínica útil para el diagnóstico de ciertas características y trastornos de la personalidad.

Entre otras pruebas de inteligencia de rango amplio se encuentran las Escalas de Habilidad Diferencial (DAS), la Prueba Detroit de Habilidad para el Aprendizaje (DTLA), la Batería de Evaluación para Niños de Kaufman (K-ABC), las Pruebas de Habilidades Cognoscitivas de Woodcock-Johnson III (WJ III) y el Sistema de Evaluación Cognoscitiva de Das-Naglieri (CAS). Las últimas tres merecen especial mención por sus bases en la teoría psicométrica y neuropsicológica.

Como representantes de las múltiples pruebas de inteligencia con fines específicos que se aplican individualmente figuran las pruebas pictóricas como la Escala de Madurez Mental de Columbia y otras pruebas de tarea única como los Laberintos de Porteus y los Diseños de Cubos de Kohs. Las pruebas de inteligencia no verbales de aplicación individual, diseñadas para personas con desventajas lingüísticas o físicas incluyen las baterías de pruebas de ejecución como las Pruebas Hiskey-Nebraska de Habilidad de Aprendizaje, la Escala Leiter de Desempeño Internacional, la Prueba Comprensiva de Inteligencia No Verbal y la Prueba Universal de Inteligencia No Verbal.

De uso más extenso que las pruebas de inteligencia individuales en escuelas y otras organizaciones, son las pruebas de inteligencia de aplicación colectiva. Estas pruebas provienen de los Exámenes Army Alfa y Army Beta del Ejército, que se basan en el trabajo pionero de Arthur Otis y otros psicólogos durante la Primera Guerra Mundial. Como ejemplos de las pruebas de inteligencia colectivas actuales están disponibles también la Prueba de Habilidad Escolar de Otis-Lennon, la Prueba de Habilidades Cognoscitivas y la Prueba de Personal Wonderlic, así como las pruebas en grupo supuestamente justas para las culturas, tales como la Prueba de Dibujo de Goodenough-Harris, las Matrices Progresivas de Raven, las Pruebas de Inteligencia Justas para las Culturas y la Prueba de Habilidad No Verbal de Naglieri.

Algunas pruebas de habilidad académica, como la Prueba de Evaluación Académica (SAT), las Pruebas Universitarias Estadounidenses (ACT) y los Exámenes del Registro de Graduados (GRE), en particular, se usan ampliamente para el ingreso a universidades e instituciones profesionales.

PREGUNTAS Y ACTIVIDADES

1. Elija una de las teorías sobre la inteligencia examinadas en este capítulo para efectuar un análisis más profundo y escriba un breve ensayo sobre su investigación.

2. ¿Cuál es la razón CI para un niño de 8 años 9 meses si su calificación en la Escala de Inteligencia Stanford-Binet es igual a la edad mental de 6 años 5 meses?

3. ¿Por qué las calificaciones del CI de desviación se consideran psicométricamente superiores a las de la razón CI?

4. Explique el desarrollo de la Escala de Inteligencia de Stanford-Binet desde las primeras pruebas de Binet hasta la cuarta edición de la escala.

5. Enumere y describa las ediciones actuales de la serie de pruebas de inteligencia de Wechsler, incluyendo el rango de edad apropiado para cada una así como las subpruebas que abarcan.

6. Compare las escalas Wechsler con las ediciones anteriores y recientes de la escala Stanford-Binet en términos de rango de edad, tipos de habilidades medidas, justicia de las pruebas para las personas con desventajas físicas o culturales, y otras características relevantes.

7. ¿Qué pruebas de inteligencia recomendaría para administrar en cada uno de los siguientes individuos? (a) Un niño de 5 años de edad en quien se sospecha retraso mental; (b) un grupo de aborígenes de las Islas del Sur; (c) un niño de 10 años con parálisis cerebral; (d) un adulto normal angloparlante; (e) un niño de 7 años totalmente ciego; (f) un adulto esquizofrénico, y (g) un grupo de alumnos de la escuela elemental con desventajas culturales.

8. Elija una de las siguientes categorías de pruebas de inteligencia analizadas en este capítulo y un instrumento publicado que sea representativo de esa categoría: pruebas pictóricas individuales; escalas de desarrollo para niños pequeños; pruebas de inteligencia colectivas grupales de multinivel; pruebas de inteligencia colectivas grupales no verbales. Obtenga tanta información como le sea posible sobre dos pruebas de los libros de texto sobre evaluación *The Mental Measurements Yearbooks, Tests, Test Critiques* y otras fuentes (consulte *The Psychological Abstracts* y *Education Index* en particular). Escriba una reseña comparativa de esas pruebas centrándose en el diseño y el formato, los procedimientos de administración y calificación, las normas, la confiabilidad, la validez y la investigación relacionada (vea la pregunta 8 de la sección de Preguntas y Actividades en el capítulo 6). Deduzca conclusiones apropiadas en cuanto a los méritos relativos de las dos pruebas que haya seleccionado.

9. ¿Cuál es la diferencia entre una prueba independiente de la cultura y otra justa para las culturas? ¿Es posible desarrollar una prueba de alguno de estos dos tipos y, de ser así, para qué se usaría?

CAPÍTULO OCHO

DIFERENCIAS INDIVIDUALES Y DE GRUPO EN LAS HABILIDADES MENTALES

Este capítulo hace una pausa en la descripción de la multiplicidad de instrumentos de evaluación cognoscitiva y se concentra más bien en tratar acerca de las diferencias individuales y de grupo detectadas en las habilidades mentales. La investigación y las aplicaciones que conciernen a las diferencias en las habilidades humanas caen en el terreno de la *picología diferencial*. El presente capítulo se limita a la descripción e interpretación de los hallazgos empíricos que conciernen a las diferencias en inteligencia y variables cognoscitivas relacionadas. Los lectores que estén interesados en un tratamiento más comprensivo de la psicología diferencial encontrarán un buen punto de partida en el libro del propio autor de la presente obra *Human Differences* (Aiken, 1999).

RETRASO MENTAL, SUPERDOTADOS Y CREATIVIDAD

Los niños y los adultos que tienen habilidades muy altas o muy bajas son de particular interés para los psicólogos y educadores preocupados por identificar a individuos situados en todos los niveles de habilidad —y por diseñar programas para tratarlos, entrenarlos y educarlos. Por supuesto, las diferencias entre las habilidades de un solo individuo pueden ser tan grandes como las detectadas entre un grupo. Por ejemplo, los niños que obtienen puntuaciones muy altas o muy bajas en las pruebas de inteligencia no por fuerza califican alto o bajo en cada medida de una habilidad cognoscitiva específica. Una persona puede ser buena en una habilidad cognoscitiva, deficiente en otra y promedio en otras más. En cualquier caso, se ha dedicado mucha atención profesional y popular a los individuos que obtienen puntuaciones muy bajas o muy altas en las pruebas de habilidad mental general. Se trata de personas retrasadas mentales o superdotadas que, dependiendo de circunstancias particulares y del punto de vista, pueden ser una pesadilla o una bendición para la sociedad.

Retraso mental

La razón principal de Alfred Binet para elaborar la primera prueba práctica de inteligencia fue identificar a los niños que tenían poca oportunidad de lograr un progreso razonable en las clases

regulares de la escuela. Por ende, no es sorprendente que uno de los usos más populares de las pruebas de inteligencia general haya sido el diagnóstico del retraso mental.

Diagnóstico y clasificación. La administración de una prueba de inteligencia no es obligatoria en el diagnóstico del retraso mental, pero, por lo general, al hacer el diagnóstico se tienen en cuenta las calificaciones obtenidas en la prueba de inteligencia junto con medidas de logro académico y vocacional, habilidades psicomotrices, madurez socioemocional y otras *conductas adaptativas*. Las conductas adaptativas pueden evaluarse mediante un análisis informal de la historia de la persona y su conducta presente, o mediante la administración de un instrumento estandarizado como las Escalas de Conducta Adaptativa de Vineland (del American Guidance Service) o las Escalas de Conducta Adaptativa AAMR (de pro.ed). El examinador llena las Escalas de Conducta Adaptativa de Vineland o AAMR con la información proporcionada por un padre, maestro u otra persona que esté familiarizada con la conducta del niño.[1]

Las etiquetas socialmente despectivas como *tarado, imbécil* e *idiota*, que se emplearon en los primeros años del siglo XX para designar a los grados alto, medio y bajo de la "debilidad mental" ya no son utilizadas por los psicólogos profesionales y educadores de Estados Unidos. Al inicio de la década de 1980, la Asociación Estadounidense de la Deficiencia Mental cambió su nombre por el de Asociación Estadounidense del Retraso Mental, y desde entonces se han hecho esfuerzos por reemplazar el término retraso mental con un término quizá menos estigmatizante como *deterioro mental* o *discapacidad del desarrollo*. En cualquier caso, se han recomendado varios sistemas para la clasificación del retraso mental que hacen uso de las calificaciones de CI. Entre éstos se encuentran los sistemas de clasificación de la Asociación Nacional para los Niños Retrasados (NARC) y la Asociación Psiquiátrica Estadounidense (APA). El sistema NARC consta de las siguientes categorías: marginalmente independiente (CI = 50 a 75), semidependiente (CI = 25 a 50) y dependiente (CI = 0 a 25). La Asociación Psiquiátrica Estadounidense (1994) menciona tres requisitos para diagnosticar el retraso mental:

1. Un funcionamiento intelectual significativamente inferior al promedio; un CI aproximado de 70 o menos en una prueba de CI administrada de manera individual (para los infantes, un juicio clínico de funcionamiento intelectual significativamente inferior al promedio).
2. Déficits o deterioros concurrentes en el comportamiento adaptativo presente (es decir, la efectividad de la persona para cumplir los estándares que su grupo cultural espera para su edad) al menos en dos de las siguientes áreas: comunicación, autocuidado, vida en el hogar, habilidades sociales/interpersonales, uso de recursos de la comunidad, autodirección, habilidades académicas funcionales, trabajo, tiempo libre, salud y seguridad.
3. El inicio es antes de los 18 años. (p. 50).

Los cuatro niveles de severidad en el sistema de clasificación del retraso mental de la APA son retraso mental leve (nivel CI de 50-55 a aproximadamente 70); retraso mental moderado (nivel CI de 35-40 a 50-55); retraso mental severo (nivel CI de 20-25 a 35-40); retraso mental profundo (nivel CI por debajo de 20 o 25). Los individuos con retraso mental leve requieren apoyo intermitente, los de retraso moderado requieren apoyo limitado, los de retraso severo requieren

[1]El reconocimiento de la importancia de los logros académicos y vocacionales, las habilidades motrices, la madurez socioemocional y otros indicadores del funcionamiento independiente, y el mantenimiento y la habilidad para cumplir las demandas culturales en cuanto a la conducta personal y social, ha llevado a la inclusión de la conducta adaptativa en el diagnóstico y clasificación del retraso mental.

considerable apoyo, y los que presentan retraso profundo necesitan apoyo sólido en sus actividades constructivas y funcionamiento social.

En las escuelas se usa en ocasiones otro sistema de clasificación que enfatiza la conducta adaptativa en lugar de la deficiencia mental: *deterioro mental susceptible de recibir educación*, para niños que tienen retraso leve; *deterioro mental susceptible a recibir capacitación*, para niños con retraso moderado; *entrenable* (*dependiente*), para niños con retraso severo, y *custodial* (*apoyo para la vida*), para los que tienen retraso profundo (Sattler, 1988).

La definición del retraso mental propuesta por la Asociación Estadounidense para el Retraso Mental (1992) se basa menos en el concepto de CI. Esta definición describe el retraso mental en términos de limitaciones sustanciales en el funcionamiento manifiesto caracterizadas por un funcionamiento intelectual significativamente inferior al promedio, el cual se presenta antes de los 18 años, y que existe de manera concurrente con limitaciones relacionadas en dos o más de las siguientes áreas de habilidades adaptativas: comunicación, vida en el hogar, uso de la comunidad, salud y seguridad, tiempo libre, autocuidado, habilidades sociales, autodirección, funcionalidad académica y trabajo. Sin embargo, esta definición ha sido criticada por algunos profesionales (por ejemplo, Jacobson y Mullick, 1992) y no se ha usado de manera amplia.

En la tabla 8.1 se presentan descripciones de las conductas características de los niños en las cuatro categorías designadas por la Asociación Psiquiátrica Estadounidense para tres periodos del desarrollo. Como se describe en esa tabla, las conductas esperadas varían con el grado de retraso y la edad cronológica del individuo. Por supuesto, esas conductas son normas o promedios, y el grado en que la conducta de un individuo en particular corresponda a las normas varía con sus antecedentes socioculturales, otras habilidades o características, y circunstancias adicionales.

Incidencia y causas del retraso. Se estima que entre 2 y $2\frac{1}{2}\%$ de la población de Estados Unidos sufre retraso mental, con un porcentaje menor de mujeres que de hombres.[2] El número de retrasados mentales es mayor entre los blancos que en todos los otros grupos étnicos, pero el porcentaje de escolares negros identificados como retrasados mentales es más alto que para cualquier otro grupo étnico, seguido de los indígenas estadounidenses, los blancos, los hispanos, y los asiáticos/isleños del Pacífico, en ese orden (U.S. Department of Education, 1997).

Si bien tanto los factores genéticos como los ambientales participan en su etiología, en tres cuartas partes de los casos se desconoce la causa exacta del retraso mental (Zigler y Hodapp, 1986). En Estados Unidos el retraso mental leve está asociado con una serie de variables demográficas relacionadas con la baja posición socioeconómica: nivel educativo bajo, pertenencia a un grupo minoritario, desempleo o niveles bajos de empleo, mala nutrición, mala salud, y condiciones de vida que suelen estar por debajo del estándar. También contribuyen al grado de retraso mental el descuido, los bajos niveles de estimulación intelectual, la escasez de experiencias formales de aprendizaje, modelos inadecuados de lenguaje y los ambientes no estructurados e impredecibles en que viven muchos niños.

El CI de los niños con retraso mental que al parecer están libres de patología orgánica, por lo común está mucho más cerca del CI promedio de la población general que el de quienes pa-

[2]Sin embargo, el porcentaje exacto varía con la calificación CI límite y las pruebas y normas específicas a partir de las cuales se determinó. Flynn (2000) afirma que, debido a que el criterio CI de 70 para un diagnóstico de retraso mental ha cambiado de las normas basadas sólo en los blancos a normas basadas en todos los estadounidenses, la proporción de individuos a quienes puede clasificarse como retrasados mentales ha fluctuado de una alta de 1 en 23 a una baja de 1 en 213 durante los pasados 50 años.

TABLA 8.1 Cambios conductuales relacionados con la edad en las personas con retraso mental

RETRASO MENTAL LEVE (CI = 50-70)
Edad preescolar (*0-5*): más lentos que el promedio para caminar, comer por sí mismos y hablar, pero el observador casual puede no advertir el retraso.
Edad escolar (*6-21*): aprenden las habilidades perceptuales-motrices y cognoscitivas (lectura y aritmética) en niveles de tercero a sexto grado al final de la adolescencia; pueden aprender a adaptarse a la sociedad.
Adulto (*21 en adelante*): por lo general alcanzan las habilidades sociales y vocacionales que necesitan para cuidar de sí mismos; requieren orientación y ayuda cuando se encuentran bajo estrés económico o social inusual.
RETRASO MENTAL MODERADO (CI = 35-49)
Edad preescolar (*0-5*): retraso perceptible en la mayor parte del desarrollo, sobre todo en el habla; pueden ser entrenados en una variedad de actividades de autoayuda.
Edad escolar (*6-21*): aprenden a comunicarse y a encargarse de las necesidades elementales de salud y seguridad; aprenden habilidades manuales sencillas, pero logran poco o ningún progreso en lectura y aritmética.
Adulto (*21 en adelante*): bajo condiciones de supervisión, realizan tareas sencillas que requieren poca o ninguna habilidad; participan en juegos sencillos y se trasladan solos en lugares familiares; son incapaces de lograr su propia manutención.
RETRASO MENTAL SEVERO (CI = 20-34)
Edad preescolar (*0-5*): demora pronunciada en el desarrollo motriz; no hablan o hablan poco; se benefician del entrenamiento en autoayuda (por ejemplo, a comer por sí mismos).
Edad escolar (*6-21*): por lo general caminan a menos que esté presente una discapacidad psicomotriz; pueden entender y responden al habla; pueden beneficiarse del entrenamiento en hábitos de salud y otros hábitos aceptables.
Adultos (*21 en adelante*): siguen rutinas diarias y contribuyen a su cuidado; necesitan dirección y supervisión cercanas en un ambiente controlado.
RETRASO MENTAL PROFUNDO (CI INFERIOR A 20)
Edad preescolar (*0-5*): retrasos extremos en todas las áreas; habilidades sensoriomotrices mínimas; requiere cuidado de una enfermera.
Edad escolar (*6-21*): es obvio que están demorados en todas las áreas del desarrollo; responden con emociones básicas y pueden beneficiarse del entrenamiento del uso de las extremidades y la boca; requieren de supervisión cercana.
Adulto (*21 en adelante*): pueden ser capaces de caminar y hablar de manera primitiva; se benefician de la actividad física regular; no pueden cuidarse por sí mismos y requieren del cuidado de una enfermera.

decen trastornos orgánicos demostrables, es decir, en el rango del retraso leve. El retraso extremo de las personas que caen en las categorías severa y profunda, y en algunos casos en la categoría moderada, se debe a una variedad de trastornos que llevan al daño del sistema nervioso central: problemas genéticos importantes como galactosemia, gargolismo, fenilcetonuria y la enfermedad de Tay-Sachs; condiciones dependientes de la genética como cretinismo, hidrocefalia y microcefalia; anormalidades cromosómicas como el síndrome de Down y el síndrome de

Klinefelter; infecciones intrauterinas; trauma del nacimiento (lesiones en la cabeza, privación o exceso de oxígeno); y enfermedades contraídas durante la infancia (meningitis, encefalitis, envenenamiento con plomo, y otras). La causa genética más común del retraso mental es el síndrome de Down, y la segunda causa más común es el síndrome del X frágil. En muchos casos, los niños retrasados en los que se desconoce la base orgánica para su condición resultan tener el síndrome del X frágil (Dykens, Hodapp y Leckman, 1994). Es probable que la investigación futura revele otras causas genéticas del retraso mental.

Los factores biológicos también pueden desempeñar un papel en las diferencias culturales en el retraso mental. Por ejemplo, las condiciones mencionadas en el párrafo anterior explican un porcentaje relativamente pequeño del número total de niños retrasados en los países más desarrollados, donde el cuidado adecuado de la salud de la madre y el infante es la regla. En los países menos desarrollados, donde la desnutrición es más común y el cuidado de la salud menos adecuado, los trastornos de desnutrición explican una alta proporción de los casos de retraso mental.

Tratamiento del retraso mental. En ocasiones el retraso mental puede recibir tratamiento médico cuando la causa se identifica de manera oportuna. Sin embargo, en la mayoría de los casos la condición es incurable, y se prescriben entrenamiento y educación en lugar de tratamientos físicos o químicos. Las personas retrasadas que reciben apoyos educacionales y sociales apropiados a lo largo de un periodo sostenido por lo general mejoran. Ese cuidado se proporciona principalmente en el hogar, aunque también se dispone de instalaciones residenciales privadas y de instituciones operadas por el Estado. En Estados Unidos la educación especial para los retrasados mentales y otros niños discapacitados es un mandato legal (Acta de Educación para todos los Niños Discapacitados, P.L. 94-142) y está disponible en todo el país. Sin embargo, dicha educación no es verdaderamente "especial", más bien consiste en procedimientos instruccionales estándar combinados con una mayor atención a las necesidades de los estudiantes. Proyectos de alcance nacional como el Proyecto Abecedarian (Campbell y Ramey, 1994; Ramey *et al*., 2000), el Proyecto Ypsilanti (Schweinhart y Weikart, 1997) y el Head Start se basaron en la idea de modificar y mejorar el desarrollo intelectual y social (Zigler, 1988). Dichos programas de intervención produjeron una eficiencia algo mayor en el aprendizaje y la adaptación social, pero las ganancias a largo plazo en las habilidades cognoscitivas fueron mínimas (Robinson, Zigler y Gallagher, 2000).

Superdotados

En el otro extremo del continuo de inteligencia del retraso mental se encuentran los superdotados. El estudio longitudinal más comprensivo de personas con CI elevado fue conducido por Lewis Terman y sus asociados (Terman y Oden, 1959). Varios cientos de niños que calificaron en el 1% superior de la distribución de CI en la Escala de Inteligencia de Stanford-Binet fueron seguidos a lo largo de sus vidas a intervalos de cinco años a partir de 1921. Después de la muerte de Terman en 1956, el estudio fue continuado por M. H. Oden (1968) y Robert Sears (1977). El propósito del estudio era obtener información sobre el éxito ocupacional, la salud física y mental, la adaptación social y otras variables asociadas con la inteligencia elevada. A partir de cuestionarios se obtuvieron detalles de la niñez, educación, personalidad, carrera(s), familia, salud física y mental, tensiones vitales de los participantes y sobre su adaptación a la vejez.

Características de los niños de Terman. Los resultados del estudio de Terman parecen contradecir una serie de mitos populares concernientes a los superdotados: que los niños brillantes son enfermizos, que se acaban pronto ("maduran pronto, se pudren pronto") y que el genio es cercano a la demencia. Esos niños mentalmente superdotados, o "termitas", eran físicamente superiores a otros niños: pesaron más al nacer y siguieron pesando más que el promedio; caminaron y hablaron más pronto y maduraron a una edad más temprana que el promedio, y su salud general era mejor. Además, cuando adultos mantuvieron su superioridad mental y física. Los datos de seguimiento revelaron que, en comparación con los adultos promedio, los superdotados obtuvieron más grados, alcanzaron mayor éxito ocupacional y mayores salarios, tenían un personal y social equivalente o mejor, lograron mayor éxito matrimonial y disfrutaban de mayor salud física. Sin embargo, el mayor éxito ocupacional de las "termitas" pareció deberse a sus mayores logros educativos más que a su CI superior *per se*. Cuando se controlaba estadísticamente el nivel educativo, las CI obtenidas en la niñez no tenían relación con el logro ocupacional. Muchas de las "termitas" no lograron vivir de acuerdo con su potencial y cuando adultos expresaron pesar por no haberlo hecho (Gardner, 1997).

Los hallazgos de Terman de una mejor adaptación y menor tasa de trastornos mentales entre los superdotados no dejaron de ser cuestionados. Hughes y Converse (1962) sugirieron que el hecho de que en principio los niños hubieran sido seleccionados sobre la base de las de los maestros, así como por el CI, puede haber sesgado la muestra a favor de los niños con buen comportamiento. Los niños superdotados de Terman también tendían a tener una posición socioeconómica por encima del promedio, lo cual también se asocia con una mejor adaptación personal.

Personalidad de los superdotados. La investigación subsecuente ha planteado también preguntas concernientes a los ajustes de personalidad de los superdotados. Webb y Meckstroth (1982) caracterizaron a los niños superdotados como más inquisitivos, activos y llenos de energía, pero también percibidos por los otros como odiosos, indisciplinados, de fuerte voluntad, traviesos, difíciles de manejar y rebeldes. Esos investigadores advirtieron que los niños superdotados a menudo son problemáticos para sus padres y se sienten atribulados. Esto parece ser más el caso de los niños enormemente talentosos con CI por encima de 150 que de niños moderadamente talentosos con CI entre 130 y 150. Los niños sumamente talentosos, por lo general, pueden leer antes de la edad para ingresar al jardín de niños y son superiores en la resolución de problemas y en otros tipos de pensamiento abstracto. Muchos se fascinan con los patrones numéricos y musicales y con la creación de nuevos enfoques y soluciones (Jackson, 1992). Pueden memorizar una partitura musical entera, averiguar cómo identificar todos los números primos o descubrir por sí mismos las reglas algebraicas (Feldman y Goldsmith, 1991; Winner, 1996).

Al igual que otros niños y adultos, los individuos superdotados son susceptibles a los trastornos psicológicos (Silverman, 1995). Al darse cuenta de que son diferentes a los otros niños, quienes son extremadamente superdotados pueden volverse independientes, inconformes, introvertidos y muy egocéntricos acerca de sus habilidades. Supuestamente conscientes de la envidia de sus compañeros de juegos y abrumados por las altas expectativas, tienden a tener una tasa más alta de problemas socioemocionales. Quienes son particularmente sensibles y están bajo gran presión para desempeñarse en público pueden deprimirse, usar drogas, no lograr desempeñarse al nivel de su habilidad y, en ocasiones, marginarse por completo de la sociedad (Janos y Robinson, 1985; Ochse, 1991).

Niños superdotados para las matemáticas. Se han conducido muchas investigaciones de niños con habilidades especiales altamente desarrolladas. Por ejemplo, Julian Stanley y sus coinvestigadores (Keating, 1976; Stanley, Keating y Fox, 1974) condujeron una serie de estudios de

preadolescentes que obtuvieron calificaciones estándar de 700 y superiores en la Prueba de Aptitud Escolar-Matemáticas (SAT-M). Los niños fueron sometidos a varias pruebas psicológicas y supervisados mientras participaban en cursos universitarios de matemáticas. Como sucede con otros niños superdotados, los niños con talento para las matemáticas a menudo aprenden asuntos complejos sin que se les enseñen de manera explícita. Los investigadores encontraron que esos niños no sólo se benefician de la instrucción a nivel universitario en matemáticas, sino que, a pesar de las preocupaciones iniciales de que pudieran ser incapaces de adaptarse al ambiente universitario, la mayoría de ellos de hecho se adaptó bien. A diferencia de otros hallazgos que conciernen a las personas superdotadas y creativas, los adolescentes con talento para las matemáticas —en el estudio de Stanley— tendieron a mostrar buena adaptación personal y alta motivación (sobre todo en matemáticas).

Educación de los niños superdotados y talentosos. Los maestros y el personal administrativo escolar utilizan el término "superdotados y talentosos" para designar a los niños con altas habilidades intelectuales u otras habilidades cognoscitivas. Por lo general, los niños en esta categoría tienen cocientes intelectuales aproximados de 130 y más altos, pero las clasificaciones y recomendaciones de los maestros, y otros criterios, también pueden contribuir a la designación de un niño como superdotado o talentoso. De acuerdo con la Ley Pública 95-561:

> Niños superdotados y talentosos significa niños, y siempre que sea aplicable, jóvenes, a quienes se identifica al nivel de preescolar, primaria o secundaria como poseedores de habilidades demostradas o potenciales que dan evidencia de una alta capacidad de desempeño en áreas como la intelectual, creativa, académica específica o de liderazgo, o en las artes visuales o interpretativas y quienes por esa razón requieren servicios o actividades que por lo general no son proporcionadas por la escuela.[3]

De acuerdo con los datos publicados por la Oficina para los Derechos Civiles del Departamento de Educación de Estados Unidos (1997), aproximadamente 6% de los escolares estadounidenses son superdotados o talentosos. Porcentajes algo más altos de mujeres que de hombres y porcentajes mayores de asiáticos/isleños del Pacífico y blancos que de indígenas americanos, hispanos y negros son clasificados como superdotados o talentosos. Algunos son excepcionales en matemáticas, otros en razonamiento verbal, otros en música o arte y otros más en liderazgo social.

Las estrategias para educar a los niños superdotados y talentosos incluyen la admisión temprana a la escuela, aceleración y salto de grados, estudio avanzado, estudio independiente, uso de mentores, enriquecimiento, clases especiales, recintos con recursos especiales y escuelas especiales. En la actualidad casi todos los sistemas escolares en Estados Unidos tienen algún tipo de programa instruccional especial para los niños superdotados. Los estudiantes inscritos en esos programas pasan la mayor parte de su tiempo escolar en las aulas regulares, pero cada semana son sacados de clase para participar en actividades especiales para los superdotados. A lo largo de Estados Unidos también se han establecido centros regionales para los niños superdotados y talentosos, así como otras instituciones dedicadas a los estudiantes con habilidades superiores. En general, a los estudiantes superdotados les va bien en lo intelectual, social y emocional en esos programas. Sin embargo, los críticos a menudo caracterizan los programas especiales para los superdotados como elitistas o antidemocráticos y recomiendan que sean suspendidos.

[3]*Congressional Record,* 10 de octubre de 1978. Enmiendas educativas de 1978, 20 USC 2701 (1978); 92 STAT.2143.

Creatividad

Las pruebas de inteligencia o de aptitud escolar administradas a los niños de edad escolar, por lo general, dan buenos resultados en la predicción del aprovechamiento escolar a corto plazo y criterios relacionados. Sin embargo, esas pruebas no fueron diseñadas para medir variables situacionales, determinación de toda la vida, motivación o talento no escolar del tipo que influye en el desempeño creativo. Llama la atención que pocos, si es que hubo alguno, de los individuos intelectualmente superdotados estudiados por Terman (Terman y Oden, 1959) alcanzaron la eminencia de un Winston Churchill, un Albert Einstein o un Ernest Hemingway. Ninguno de ellos se convirtió tampoco en un compositor, artista o poeta famoso.

Características de la gente creativa. Thomas Alva Edison poseía 1093 patentes, Albert Einstein publicó 248 trabajos, Pablo Picasso promedió más de 200 obras de arte en un año, y Wolfgang Amadeus Mozart compuso más de 609 piezas musicales durante su corta vida; murió a los 35 años. Esos casos ilustran la elevada pulsión interna que poseen muchas personas creativas (Haney, 1985). Otros rasgos afectivos y cognoscitivos que se dice caracterizan a las personas creativas son la fluidez de ideas, la flexibilidad, la falta de convencionalismos, la sensibilidad social, no estar a la defensiva, una mayor voluntad para concederse fallos y vínculos cercanos con los padres (MacKinnon, 1962).

De acuerdo con los resultados de las investigaciones de MacKinnon (1962) y Wallach y Kogan (1965), parecería que la creatividad, en especial cuando se acompaña por una inteligencia elevada, no es una mala característica desde el punto de vista de la salud mental. Sin embargo, en un estudio de artistas británicos destacados (novelistas, pintores, dramaturgos, poetas y escultores), Jamison (1989, 1993) encontró que esos individuos tenían una probabilidad mucho mayor que la gente menos creativa de haber sido tratados por trastornos del estado de ánimo (manía y depresión). Andreasen (1987) encontró resultados similares en un estudio de 30 miembros del cuerpo docente en un taller para escritores: 80% exhibió depresión o alguna otra forma de trastorno del estado de ánimo y a 43% se le diagnosticó como maniaco-depresivos. El significado de esos hallazgos no es del todo claro, pero al menos sugiere que los adultos creativos, como los niños superdotados, no desconocen la infelicidad y la mala adaptación (vea también Ludwig, 1995).

Pruebas de creatividad. En ocasiones se afirma que la inteligencia por arriba del promedio es necesaria pero no suficiente para la productividad creativa. Más allá de un nivel mínimo de inteligencia, el desempeño creativo parece depender más de la motivación y las habilidades especiales que de la habilidad mental general (MacKinnon, 1962). Por consiguiente, las investigaciones de la creatividad conducidas durante los pasados 40 años se han concentrado en identificar otras características cognoscitivas y afectivas que distinguen a la gente creativa de la no creativa. Por ejemplo, se han hecho esfuerzos por desarrollar medidas de la habilidad de pensamiento divergente en oposición al convergente (Guilford, 1967). En las medidas de *pensamiento convergente*, como los problemas del tipo que se encuentra en las pruebas de inteligencia, hay una sola respuesta correcta. En contraste, en las pruebas de *pensamiento divergente*, a los sujetos se les presentan problemas flexibles que tienen varias soluciones posibles y se califica la originalidad de sus respuestas. Por desgracia, esta flexibilidad crea dificultades en la calificación y en la determinación de la confiabilidad y la validez de esas pruebas. Entre los procedimientos de calificación propuestos está la evaluación de acuerdo con el número de respuestas dadas por el examinado (*fluidez*) y su originalidad o singularidad (*novedad*).

Los siguientes son ejemplos de reactivos de pruebas de creatividad:

Prueba de consecuencias. Imagine todas las cosas que podrían suceder si, de repente, se abolieran todas las leyes nacionales y regionales (Guilford, 1954).

Prueba de asociaciones remotas. Encuentre una cuarta palabra que se asocie con cada una de estas tres palabras: (a) rata-azul-casita, (b) fuera-perro-gato, (c) rueda-eléctrico-alto, (d) sorpresa-línea-cumpleaños (Mednick, 1962).

Prueba de usos poco comunes. Mencione tantos usos como pueda pensar para (a) un mondadientes, (b) un ladrillo y (c) un clip para papel (Guilford, 1954).

Prueba de asociación de palabras. Escriba tantos significados como pueda para cada una de las siguientes palabras: (a) pato, (b) costal, (c) resina y (d) justo (Getzels y Jackson, 1962; copyright © 1962, John Wiley & Sons, Inc. Reproducido con autorización de John Wiley & Sons, Inc.).

Las baterías de pruebas de creatividad, como las Pruebas de la Estructura del Intelecto (de Consulting Psychologists Press) y las Pruebas Torrance de Pensamiento Creativo (TTCT) (de Scholastic Testing Service), representan una combinación de medidas de creatividad. La TTCT consta de tres ejercicios basados en ilustraciones (TTCT Figurativo: Pensamiento Creativo con ilustración) y seis ejercicios basados en palabras (TTCT Verbal: Pensamiento Creativo con Palabras). Un ejemplo de los tipos de reactivos en la TTCT verbal es "Escriba todas las preguntas en las que pueda pensar" acerca de una determinada ilustración. En una parte de la TTCT figurativa se pide al examinado que elabore un guión a partir de una línea básica. La TTCT verbal, cuya solución se lleva 45 minutos, se califica en tres variables: fluidez, flexibilidad y originalidad. La TTCT figurativa, cuya terminación requiere 30 minutos, se califica en cinco variables: fluidez, originalidad, elaboración, abstracción de los títulos y resistencia al cierre prematuro. La TTCT se reestandarizó en 1980, y en el manual se proporcionan los rangos percentilares nacionales y las estándar desde el primer grado hasta los niveles universitario y adulto. Aunque una serie de investigaciones concluyó que la TTCT es un indicador no sesgado de la genialidad (por ejemplo, Esquivel y Lopez, 1988; Torrance, 1988), las confiabilidades de las pruebas varían mucho y los resultados de los estudios de validez no son concluyentes (Hattie, 1980).

Evaluación de las pruebas de creatividad. Las pruebas que han sido diseñadas para evaluar la creatividad son fascinantes, pero es importante considerar las críticas hechas por McNemar (1964) y otros psicólogos. Las pruebas de creatividad con frecuencia tienen correlaciones significativas con las pruebas de CI, y al parecer las primeras no son más efectivas que las últimas para predecir el desempeño creativo. Considerando todas las cosas, una conclusión razonable es que todavía queda por demostrar si es posible construir medidas efectivas de la creatividad. Hasta que se diseñe una prueba que haga una predicción precisa del desempeño en un criterio de creatividad de aceptación general, sería conveniente seguir el consejo de McNemar (1964) de no deshacernos de nuestras pruebas de inteligencia general.

INVESTIGACIÓN SOBRE LOS CORRELATOS DEMOGRÁFICOS DE LAS HABILIDADES MENTALES

Desde el momento de su aparición en la primera década del siglo XX, las pruebas de inteligencia han formado parte de numerosas investigaciones interesadas en las características, causas y efectos de las diferencias individuales en las habilidades cognoscitivas. Por desgracia, esas investigaciones, las cuales fueron iniciadas por Francis Galton en la última parte del siglo XIX, con

mucha frecuencia han sido asistemáticas y son reflejo de métodos correlacionales convenientes en lugar de un diseño de investigación sólido. Aunque los resultados de dichos estudios pueden ser difíciles de interpretar, provocan la reflexión y deben ser tomados en cuenta por cualquiera que decida teorizar acerca de la naturaleza y el desarrollo de la cognición humana.

Diferencias de edad en las habilidades mentales

Debido a que la confiabilidad de todas las pruebas de inteligencia es menos que perfecta, la calificación de una persona en una prueba particular cambiará algo de acuerdo con el momento y las condiciones de la examinación. No obstante, dada una situación de vida relativamente estable y condiciones óptimas de examinación, las calificaciones en las pruebas de inteligencia son bastante estables en los años escolares. Las calificaciones tienden a ser menos estables en la niñez temprana y media, pero son más consistentes durante la adolescencia. El CI de un niño en una prueba individual de inteligencia varía unos cinco puntos en promedio, y los cambios de 20 puntos o más son raros. Las fluctuaciones grandes en el CI, por lo general, pueden rastrearse hasta variaciones bastante considerables en la salud o las condiciones de vida, así como remitir a problemas y experiencias emocionales graves.

La antigua definición del cociente de inteligencia como 100 veces la razón entre la edad mental y la edad cronológica implica que, para que el CI permanezca estable de un año a otro, la edad mental debe cambiar de manera proporcional a la edad cronológica. La misma suposición se aplica a las pruebas que no arrojan CI de razón: en las pruebas de inteligencia las puntuaciones crudas y la edad mental deben aumentar con la edad durante la niñez. La forma exacta de la función que relaciona las puntuaciones crudas de la prueba o edad mental con la edad cronológica depende, por supuesto, de la prueba específica y de los componentes intelectuales que ésta mide.

Estudios transversales y longitudinales. Las conclusiones de los primeros estudios de los cambios con la edad en la inteligencia que por lo general están basados en datos transversales (Doppelt y Wallace, 1955; Jones y Conrad, 1933; Yerkes, 1921). En un análisis de las calificaciones en el *Examen Army Alfa* aplicado a soldados estadounidenses durante la Primera Guerra Mundial, Yerkes (1921) encontró que las calificaciones promedio en la prueba declinaban de manera estable de finales de la adolescencia hasta la sexta década de la vida. En otro estudio temprano, Jones y Conrad (1933) encontraron que las puntuaciones promedio del Examen Alfa del ejército aumentaban linealmente de los 10 a los 16 años, pero luego declinaban gradualmente hasta llegar al nivel de los 14 años a la edad de 55 años. Las normas de la Escala de Inteligencia para Adultos de Wechsler también indicaban que la inteligencia alcanza su punto máximo en la juventud, aunque a una edad algo mayor de lo que se encontró en los primeros estudios. Las puntuaciones promedio de la escala completa en la WAIS-R alcanzan su punto máximo al principio de los 20 años, permanecen bastante constantes desde ese punto hasta finales de los 20 o principios de los 30 años, y luego declinan de manera estable a lo largo de la vida posterior.

En contraste con los *estudios longitudinales*, que comparan el desempeño del mismo grupo de personas en diferentes edades, los *estudios transversales* comparan el desempeño de grupos de personas (cortes) que crecieron bajo circunstancias ambientales diferentes. Las diferencias entre las cortes en factores como la oportunidad de educación, la cual mantiene una relación estrecha con las calificaciones en las pruebas de inteligencia, hacen difícil igualar a personas de diferentes edades. En consecuencia, es imposible comparar los niveles de inteligencia de personas de edades distintas sin confundir los efectos de la educación con los de otras experiencias relacionadas con la prueba.

El aumento estable en los niveles educativo y socioeconómico promedio de los estadounidenses durante el siglo XX debe tomarse en consideración al interpretar la declinación aparente con la edad en las habilidades cognoscitivas. Debido a que las calificaciones en las pruebas de inteligencia tienen una relación positiva con el nivel educativo y la posición socioeconómica, los adultos mayores, quienes tuvieron menos educación formal y una posición socioeconómica por lo general más baja, tienden a obtener puntuaciones de prueba significativamente más bajas que los adultos más jóvenes.

Dado que los estudios longitudinales de inteligencia han sido realizados más a menudo con graduados universitarios y otros grupos favorecidos en lo intelectual, puede argumentarse que los hallazgos no por fuerza se aplican a la población general (Bayley y Oden, 1955; Campbell, 1965; Nisbet, 1957; Owens, 1953, 1966). Sin embargo, las investigaciones longitudinales conducidas en personas de inteligencia promedio (Charles y James, 1964; Eisdorfer, 1963; Tuddenham, Blumenkrantz y Wilkin, 1968) y en adultos con retraso mental no institucionalizados (Baller, Charles y Miller, 1967; Bell y Zubek, 1960) han arrojado resultados similares. Las calificaciones promedio en las pruebas de inteligencia se incrementan en pequeñas cantidades durante la adultez temprana y se estabilizan entre los 25 y 30 años. La inteligencia de las personas que están por debajo del promedio o que no hacen un uso adecuado de sus habilidades declina un poco durante la adultez temprana. Por otro lado, los individuos de inteligencia por arriba del promedio pueden no mostrar declinación o incluso continuar mejorando bien avanzada la edad madura. Aunque los resultados de los estudios transversales y longitudinales revelan disminuciones sustanciales en las habilidades cognoscitivas durante la octava y la novena décadas, se ha encontrado que dichas habilidades pueden incrementarse incluso después de los 70 años (Baltes y Schaie, 1974; Busse y Maddox, 1985; Schaie y Hertzog, 1983). Se ha interpretado que esos estudios indican que la magnitud de la disminución intelectual con el envejecimiento varía tanto con la naturaleza de la tarea de la prueba como con el individuo.

Habilidades específicas. Las pruebas de inteligencia general miden una combinación de varias habilidades cognoscitivas, y el patrón de cambio en el desempeño con la edad varía según la habilidad específica. Como se ve en el patrón relacionado con la edad de las calificaciones escaladas del subtest en el WAIS-R (Wechsler, 1981), las calificaciones en las pruebas de vocabulario e información por lo general no muestran cambios apreciables con el envejecimiento, pero las habilidades perceptual-integrativa y de comprensión de símbolos numéricos declinan con mayor rapidez.

Tanto los métodos transversales como los longitudinales tienen desventajas y se requieren investigaciones que combinen los dos enfoques para alcanzar conclusiones válidas acerca del crecimiento intelectual con la edad. En los Estudios Longitudinales de Seattle, Schaie (1990, 1994) y sus colaboradores condujeron una serie de estudios transversales y longitudinales para analizar cambios con la edad en cinco habilidades medidas por las Pruebas de Habilidades Mentales SRA: significado verbal, orientación espacial, razonamiento inductivo, número, y fluidez de palabra. Los hallazgos demostraron que la naturaleza de la relación entre la calificación obtenida en la prueba y la edad cronológica variaba con la habilidad específica y la metodología de investigación. Sin embargo, los resultados globales demostraron que durante la madurez la tasa de declinación era mayor para orientación espacial y razonamiento inductivo y menor para fluidez de palabra, significado verbal y número. Durante la vejez la mayor caída fue en las calificaciones de significado verbal, una prueba ligeramente acelerada. Otros investigadores han encontrado una mayor declinación relacionada con la edad en la habilidad para razonar y resolver problemas que impliquen estímulos visuales y geométricos (*inteligencia fluida)* que en las habilidades verbales (*inteligencia cristalizada*) (Christensen *et al*., 1994; Horn, 1982; Horn y Hofer, 1992).

Schaie y sus coinvestigadores (Baltes y Willis, 1982; Schaie y Willis, 1986; Willis, 1990) concluyeron que las habilidades cognoscitivas muestran cierto deterioro con el envejecimiento, pero enfatizaron que esas habilidades son plásticas y que el deterioro en las mismas puede ser detenido e incluso revertido. Sostienen que proporcionar oportunidades variadas para la estimulación intelectual y un estilo de vida flexible puede contribuir al mantenimiento de un nivel óptimo de funcionamiento cognoscitivo en la vejez. Como un programa de demostración, elaboraron un conjunto de procedimientos de entrenamiento para que los adultos mayores mejoraran sus calificaciones en las pruebas de inteligencia. Dicho entrenamiento implicaba no sólo instrucción en habilidades cognoscitivas específicas, sino también reducción de la ansiedad y motivación. También se alentó a los participantes en las sesiones de entrenamiento a compensar la disminución que percibieran en ciertas habilidades cognoscitivas concentrándose menos en esas habilidades y más en las que sus déficit cognoscitivos fueran menos pronunciados.

En resumen, el hecho de que se observe con la edad una disminución, ningún cambio o incluso un incremento en las habilidades cognoscitivas depende no sólo de la metodología de investigación (longitudinal, transversal o de variaciones en esos métodos), sino también de la habilidad específica y de la persona probada. Las variaciones en las habilidades cognoscitivas durante la adultez también dependen en cierta medida de las experiencias de la persona relacionadas con la prueba. La gente que permanece activa en lo intelectual muestra a menudo menor deterioro en las calificaciones de pruebas de inteligencia que quienes no lo hacen. E incluso cuando los adultos mayores tienen un mal desempeño en las pruebas de inteligencia, pueden poseer conocimiento y habilidades muy especializadas en áreas no cubiertas por los instrumentos. Dichas habilidades pueden ayudar a los adultos mayores a ser hasta más competentes que los adultos jóvenes al tratar con los problemas de la vida cotidiana.

Caída terminal. Una excepción aparente a la conclusión de que el deterioro en las habilidades cognoscitivas en la vejez es gradual y varía con la habilidad específica es un fenómeno conocido como *caída terminal*. Este concepto se refiere a un deterioro en el funcionamiento cognoscitivo (CI, memoria, organización cognoscitiva), el tiempo de reacción y en otras habilidades sensoriomotrices y características de personalidad como la asertividad durante los últimos meses o años de vida. Un impulso para la investigación sobre la caída terminal fue la afirmación hecha por una enfermera de un asilo en el sentido de que podía predecir qué pacientes iban a morir pronto por la simple observación de que "parecían actuar de manera diferente" (Lieberman, 1965, p. 181). Los hallazgos de la investigación subsecuente revelaron deterioros en varias áreas del funcionamiento cognoscitivo y sensoriomotriz y en la habilidad para afrontar las demandas ambientales en los pacientes que murieron en el curso de un año posterior a la prueba (Granick y Patterson, 1972; Lieberman y Coplan, 1969; Reimanis y Green, 1971, y Riegel y Riegel, 1972). Riegel y Riegel (1972) advirtieron que la caída terminal era evidente hasta cinco años antes de la muerte, pero los resultados de la investigación subsecuente indicaron que tal caída puede no comenzar hasta alrededor de dos años antes de la muerte y que sólo ocurre en ciertas habilidades (White y Cunningham, 1988).

Los estudios de hombres viejos que participaron en un estudio longitudinal del envejecimiento conducido por investigadores de la Universidad de Duke no encontraron caída terminal en pruebas de funcionamiento físico, pero las calificaciones en las pruebas de inteligencia tendían a caer de manera pronunciada unos cuantos meses o años antes de la muerte (Palmore, 1982; Palmore y Cleveland, 1976; Siegler, McCarty y Logue, 1982). Era más probable que los deterioros ocurrieran en pruebas no aceleradas como las de vocabulario, el cual al parecer es poco afectado por la edad hasta tarde en la vida, que en pruebas aceleradas de naturaleza percep-

tual o de resolución de problemas. Por otro lado, los pacientes que no mostraron dichos deterioros en el funcionamiento cognoscitivo y la conducta no murieron sino hasta después de transcurrido un periodo significativamente más largo de haber sido probados.

Efecto Flynn. Otro fenómeno que tiene que ver con los cambios en la inteligencia relacionados con la edad, pero en este caso cambios a lo largo de generaciones, es el *efecto Flynn*. A partir de un análisis sobre calificaciones CI en países desarrollados a lo largo de tres generaciones, el científico político James Flynn (1987) concluyó que el CI promedio de las personas comunes de 20 años en la década de 1980 era 15 puntos más alto que el de una persona comparable en 1940, y que continuaba creciendo en un estimado de .33 puntos de CI por año. Las diferencias generacionales en el CI promedio eran mayores en pruebas como la de Matrices Progresivas de Raven, una medida de habilidad visoespacial, que en las pruebas de Wechsler y de Stanford-Binet, las cuales son medidas de vocabulario, información general, aritmética y otros conocimientos adquiridos, así como de habilidad visoespacial. Flynn concluyó que el incremento generacional observado en las calificaciones promedio de las pruebas de inteligencia se debe más a incidencias ambientales que a factores genéticos, pero que las calificaciones no podían atribuirse sólo a mejoras en la escolaridad formal. Otros factores que posiblemente contribuyen son los mayores logros educativos de los padres, la mayor atención de los padres a los niños, el progreso en la posición socioeconómica, la mejor nutrición, la disminución de las enfermedades en la niñez y una sociedad cada vez más compleja en lo tecnológico. De acuerdo con Greenfield (1998), buena parte del incremento en el CI informado por Flynn se debe a los efectos visuales especiales proporcionados por la televisión, las computadoras, los juegos de vídeo y otros instrumentos tecnológicos. También se ha notado que en las últimas décadas han disminuido de manera notable la desnutrición severa y las deficiencias en yodo, hierro y otros nutrientes asociados con menores CI, así como con menor estatura. Lynn (1998) y Sigman y Whaley (1998) encontraron que la evidencia que vincula a la inteligencia con la mejor nutrición es convincente, pero Martorell (1998) concluyó que la mejor nutrición probablemente no es responsable del efecto Flynn. Por último, debe advertirse que, si bien las puntuaciones crudas promedio en las pruebas de CI han estado aumentando por décadas, sigue siendo controvertida la cuestión de si la inteligencia de la población en realidad está aumentando (vea Howard, 2001).

Otros correlatos de las habilidades mentales

En cientos de estudios se ha examinado la relación de las calificaciones en las pruebas de inteligencia con una multitud de variables demográficas, incluyendo el tamaño de la familia, el orden de nacimiento, la ocupación, la posición socioeconómica, la educación, la nacionalidad y la cultura. La metodología y los hallazgos de esas investigaciones constituyen parte sustancial de los temas de cursos sobre psicología diferencial.

Tamaño de la familia y orden de nacimiento. En muchos estudios se ha documentado la relación inversa entre tamaño de la familia e inteligencia (Lancer y Rim, 1984; Steelman y Doby, 1983; Wagner, Schubert y Schubert, 1985). La tendencia a que las personas mentalmente más torpes provengan de familias más grandes no se debe por completo a las diferencias socioeconómicas entre las familias grandes y pequeñas, ya que sigue siendo significativa incluso cuando se consideran dichas diferencias. La relación entre el tamaño de la familia y la inteligencia es ciertamente multicausal, pero no necesariamente bidireccional. Los padres con CI bajos tienden a tener un mayor número de hijos que el promedio, pero las familias grandes no por fuerza producen hijos con bajos CI. Aunque puede ser razonable suponer que en las familias más grandes se concede menos atención al desarrollo cognoscitivo de los hijos, esto no por fuerza es cierto en la sociedad estadounidense moderna (Rodgers, Cleveland, van den Oord y Rowe, 2000).

Desde la época de Francis Galton se ha observado que los primogénitos tienen mayor probabilidad de alcanzar grandes logros que los hijos nacidos después. Resumiendo los resultados de estudios realizados hasta mediados de la década de 1960, Altus (1966) concluyó que los primogénitos constituyen un porcentaje mayor de la porción intelectualmente superior de la población que de la población como un todo. Los primogénitos también hablan antes y de manera más clara, aprenden a leer más pronto y son mejores en la resolución de problemas y tareas perceptuales que los nacidos más tarde. Una posible explicación de esas diferencias es que los padres por lo regular tratan a los primogénitos (en particular a los varones) de manera diferente a los niños que nacen después. Ambos padres tienden a prestar más atención y estimulación a sus hijos primogénitos, pasan más tiempo con ellos y los alientan y ayudan más para caminar, hablar, leer a la edad apropiada y en otras tareas del desarrollo (Altus, 1966; Lewis y Jaskir, 1983; MacPhee, Ramey y Yeates, 1984). El hallazgo de que la relación entre el tamaño de la familia, el orden de nacimiento y las habilidades intelectuales es más evidente en las medidas verbales que en las no verbales de habilidad es congruente con el énfasis de los padres en el desarrollo del lenguaje de esos niños (Lancer y Rim, 1984). También se ha pensado que las diferencias en el trato que dan los padres a los primogénitos y a los niños que nacen después son responsables de que los primogénitos sean más serios, responsables, estudiosos y competitivos, mientras que los nacidos más tarde son más sociables, relajados, imaginativos y atléticos.

Posición ocupacional. En una sociedad abierta y competitiva como la nuestra, es razonable esperar que las personas más inteligentes ingresen en ocupaciones que requieren habilidades cognoscitivas más altas. Del mismo modo, las personas de menor inteligencia tienden a entrar en ocupaciones para las cuales se necesita de menor habilidad. Uno de los hallazgos más citados en las pruebas mentales se relaciona con este punto: las diferencias en las calificaciones promedio de la Prueba de Clasificación General del Ejército (AGCT) de reclutas militares de la Segunda Guerra Mundial que habían sido empleados en varias ocupaciones civiles (Harrell y Harrell, 1945). Las calificaciones promedio en la AGCT calculadas en más de 70 grupos ocupacionales demostraron que los contadores, abogados e ingenieros estaban en la parte superior. Los conductores de camiones, mineros y granjeros se encontraban en la parte inferior, y los otros grupos ocupacionales estaban arreglados en el medio de una jerarquía de acuerdo con sus calificaciones promedio en la AGCT. Como era de esperar, hubo un amplio rango de calificaciones dentro de cada ocupación. Por ejemplo, algunos conductores de camiones calificaron más alto que algunos maestros, lo que prueba que los primeros no por necesidad son lo opuesto de los "chicos sabios". No obstante, los datos demuestran con claridad la importancia de la variable inteligencia en la predicción de la pertenencia a una ocupación. En general, las calificaciones de las pruebas de inteligencia hacen una predicción bastante buena del desempeño en una variedad de ocupaciones (Brody, 1992).

El papel de la educación, la cual tiene una relación significativa tanto con la inteligencia como con el estatus ocupacional, no está del todo claro en la determinación de la relación entre las dos últimas variables. Cronin, Daniels, Hurley, Kroch y Webber (1975) sostenían que la correlación entre la inteligencia y el estatus ocupacional se debe al hecho de que ambas variables están correlacionadas con los antecedentes de clase social. Concluyeron que los hogares de clase media o superior tienen mayor probabilidad que los hogares de clase baja de preparar a los niños para hacer un buen papel en las pruebas de inteligencia y en el trabajo escolar, pavimentando así el camino para que ingresen en ocupaciones de estatus superior. La secuencia causa-efecto también puede ser la siguiente: calificar alto en una prueba de inteligencia o de aptitud escolar, por lo general, es un requisito para la admisión a un buen colegio, y la graduación de un buen colegio o universidad (y/o de alguna escuela profesional en algunos casos) es un requisito para ingresar a una ocupación de mayor prestigio.

Posición socioeconómica. Uno de los hallazgos más consistentes sobre las diferencias individuales y de grupo en las características psicológicas es la correlación positiva entre el CI y la posición socioeconómica (PSE), donde la PSE se define en términos del ingreso, la educación y la ocupación de los padres. En esos estudios se han encontrado a menudo CI superiores al promedio entre los niños de las clases sociales más altas, una distinción que se mantiene tanto en las pruebas convencionales como en las pruebas justas para la cultura (Speath, 1976). Que las diferencias de clase social en la habilidad sean sobre todo el resultado de la herencia o del ambiente es tema de debate, pero generalmente se acepta que un ambiente familiar donde se brinde apoyo puede ejercer un efecto significativo sobre las habilidades cognoscitivas.

Debido a la estrecha relación entre la posición socioeconómica y el nivel educativo, es difícil concluir si las diferencias observadas en los CI se deben a diferencias en la educación o a alguna otra variable asociada con la posición socioeconómica. Los niños que califican bajo en las pruebas de inteligencia no sólo tienden a tener menos educación formal, sino que también provienen de hogares enajenados por la cultura dominante y que están bajo mayor presión económica que el promedio. En esos hogares suele emplearse como medio principal de comunicación un idioma distinto al inglés estándar y los padres no enfatizan la importancia de las habilidades académicas ni saben cómo ayudar a sus hijos a adquirirlas.

A pesar de la correlación positiva significativa entre las calificaciones en las pruebas de inteligencia y la posición socioeconómica, las dos variables están lejos de ser intercambiables. Considere, por ejemplo, los resultados de un estudio conducido por Thomas, Alexander y Eckland (1979) de las relaciones de esas variables con las notas escolares: se encontró que la correlación positiva entre CI y logro educativo seguía siendo significativa incluso cuando se controlaba de manera estadística la posición socioeconómica. Por otro lado, cuando el CI se controlaba de manera estadística, la correlación entre la posición socioeconómica y el logro educativo era ligeramente negativa. Esos hallazgos sugieren que la correlación entre el CI y las notas escolares no se debe, como creen algunos psicólogos, sobre todo a las diferencias en los antecedentes de clase social. Más bien, parece que la habilidad intelectual afecta tanto a la posición socioeconómica como al nivel educativo. Por ello, puede argumentarse que una razón por la cual los estudiantes de clase media tienen mayor probabilidad que los de posición socioeconómica baja de estar en la mitad superior de sus grupos escolares es porque poseen mayor habilidad intelectual (Thomas, Alexander y Eckland, 1979).

Residencia urbana y rural. El lugar de residencia (urbano contra rural) se relaciona con la pertenencia ocupacional, la posición socioeconómica y las calificaciones en las pruebas de inteligencia. Estudios realizados en Estados Unidos en la primera mitad del siglo XX (vea McNemar, 1942) encontraron que los niños que vivían en áreas rurales tenían CI promedio significativamente menor al de quienes vivían en áreas urbanas. Aunque la diferencia urbana-rural en las calificaciones de las pruebas de inteligencia ha persistido, no es tan pronunciada como en las generaciones previas. Debido a la televisión, al mejor acceso a las escuelas, a otras fuentes de información y estimulación intelectual y a los avances en la tecnología agrícola, en la actualidad los niños del campo están expuestos a una gama más amplia de estímulos ambientales y tienen mayores oportunidades de aprender que sus padres y sus abuelos. La mayor exposición a la cultura más amplia ha mejorado el vocabulario, el nivel de conocimiento y la conciencia intelectual general de los niños del campo. Reynolds, Chastain, Kaufman y McLean (1987) estimaron que las mejoras en los servicios de comunicación y transporte produjeron una caída de la diferencia promedio entre los niños urbanos y rurales desde 6 puntos CI hace una generación hasta alrededor de 2 puntos en la década de 1980. Además, estudios conducidos entre los venda de Sudáfrica, los malayos y chinos de Malasia y los nigerianos apoyan la conclusión de que las diferencias de grupo en el desempeño en las pruebas de inteligencia reflejan diferencias en la clase social y

la educación más que del ambiente urbano contra el rural *per se* (Cronbach y Drenth, 1972; Scribner y Cole, 1973). Lo mismo puede decirse de las diferencias en las calificaciones obtenidas en las pruebas por niños que viven en diferentes secciones de las áreas metropolitanas.

La dinámica del ambiente familiar va más allá de variables como el tamaño de la familia, el orden de nacimiento y la posición socioeconómica. El estilo de crianza, el proporcionar un ambiente familiar que ofrezca apoyo y otras medidas de tratamiento dentro del hogar son predictores todavía más importantes de las calificaciones obtenidas en las pruebas de inteligencia por los niños pequeños (Hunt, 1961; Molfese, DiLalla y Bunce, 1997). Sea como sea, no está del todo clara la magnitud de esos efectos en las calificaciones de los niños en las pruebas de inteligencia. Por ejemplo, los hallazgos de las investigaciones de Baumrind (1993), Jackson (1993) y Scarr (1992, 1993) indican que, si bien las características del hogar y de los padres tienen una relación significativa con las puntuaciones en las pruebas de inteligencia en la niñez temprana, para la adolescencia esos efectos se han vuelto muy pequeños.

Expectativas del maestro. Las habilidades cognoscitivas influyen ciertamente en el logro educativo, pero la educación también influye en la habilidad . Los efectos de la educación sobre las habilidades cognoscitivas en ocasiones son indirectos, como lo revelan los estudios de las expectativas del profesor. El sociólogo C. H. Cooley (1922) propuso la *teoría del espejo*, por la cual afirma que las personas tienden a adaptar su conducta y la forma en que se perciben a la manera en que creen ser percibidas por los demás. Algunos años después, las investigaciones surgidas de la observación de que los hallazgos de los investigadores a menudo se relacionan con sus expectativas se extendieron a la situación del salón de clases. Esas investigaciones, que con frecuencia implicaban a niños con desventajas sociales, se interesaban en la influencia de las expectativas y actitudes de los maestros sobre los cambios observados en las calificaciones en las pruebas y las conductas de los estudiantes. Un famoso, aunque algo controvertido, experimento de este tipo fue conducido por Rosenthal y Jacobson (1968) en las escuelas primarias del distrito escolar sur de San Francisco.

El propósito del experimento era determinar los efectos de decir a los maestros que ciertos alumnos mostrarían una "aceleración potencial" en su habilidad intelectual en el año escolar siguiente. En septiembre se obtuvieron calificaciones de CI verbal, de ejecución y total para todos los niños de la escuela al hacerlos presentar una prueba de inteligencia no verbal, las Pruebas de Habilidad General (TOGA). Luego, en un informe para los maestros, se etiquetó a 20% de los niños como "aceleradores potenciales", supuestamente sobre la base de sus calificaciones en la TOGA, pero en realidad se hizo al azar. La TOGA volvió a administrarse a todos los niños un semestre, un año y dos años más tarde. Se hicieron entonces comparaciones entre las ganancias en el CI de los grupos experimentales ("aceleradores potenciales") y las de los grupos control de niños que no fueron etiquetados como aceleradores potenciales. Las ganancias en el CI de los grupos experimentales de primero a tercer grado fueron significativamente mayores que las de los controles, pero las diferencias CI entre los grupos experimentales y los controles de cuarto a sexto grado no fueron significativas. Los niños de origen mexicano y los de habilidad media mostraron mayores ganancias iniciales en el CI total. Los varones mostraron ganancias promedio más grandes en el CI verbal y las niñas en el CI de razonamiento. Los niños experimentales también mostraron mayores ganancias en lectura y fueron calificados por sus maestros como más felices, intelectualmente más curiosos y menos necesitados de aprobación social que los controles.

Rosenthal y Jacobson no pudieron identificar las conductas específicas del maestro que producían los cambios en el CI para los grupos experimentales, pero especularon que las mayores expectativas de los maestros para esos niños fueron comunicadas por medio de expresiones

faciales, posturas, tacto y otras señales no verbales. Los hallazgos de este experimento no fueron replicados completamente por otros investigadores, y se le criticó por una serie de defectos metodológicos. Además, un meta-análisis subsecuente de los estudios sobre el efecto de las expectativas dio firme apoyo a la hipótesis de que entre más familiarizados estén los maestros con sus alumnos menor es el efecto de las expectativas del maestro sobre las calificaciones CI de los niños (Raudenbush, 1984).

Nacionalidad. De acuerdo con el dogma popular, ciertas nacionalidades y grupos étnicos poseen características específicas de conducta y personalidad que los distinguen de otros grupos de personas. Aunque esos estereotipos contienen un elemento de verdad, por lo regular son generalizaciones excesivas que pueden servir como justificaciones para el tratamiento diferencial o incluso para el maltrato de grupos nacionales y étnicos particulares. No obstante, los científicos sociales han mostrado un interés considerable en las relaciones de las variables cognoscitivas con la nacionalidad, el grupo étnico y la cultura.

Varias investigaciones tempranas interesadas en las diferencias de grupo que probablemente inciden en la inteligencia se concentraron en la nacionalidad. Un estudio influyente realizado en la década de 1920 concluyó que los inmigrantes judíos, escandinavos y alemanes (junto con los estadounidenses nativos) obtenían en las pruebas de inteligencia calificaciones promedio superiores a las de otros grupos de inmigrantes en Estados Unidos (Hirsch, 1926). Esos resultados, los cuales sugerían que los inmigrantes de países del norte y el occidente de Europa eran más inteligentes que los de otros países, causaron tal impresión en el psicólogo H. H. Goddard que cabildeó a favor de leyes de inmigración que restringieran la admisión a Estados Unidos de todos los inmigrantes a excepción de los del norte y el occidente de Europa (Gould, 1981). Más tarde se interpretó que los hallazgos de Hirsch (1926), combinados con los de Yerkes (1921), Brigham (1923) y otros, se debían a la migración selectiva; no se encontraron diferencias significativas de nacionalidad cuando se probó a las personas en sus países nativos y en su lengua materna. En particular, Brigham (1930) repudió sus afirmaciones concernientes a las diferencias de nacionalidad en el Examen Army Alfa, y concluyó que los métodos utilizados fueron erróneos y que las pruebas medían la familiaridad con el lenguaje y la cultura estadounidenses más que la inteligencia innata. En otros estudios de inmigrantes se encontró que las calificaciones en las pruebas estadounidenses de inteligencia variaban con la semejanza entre la cultura nativa de los examinados y la cultura estadounidense dominante.

Ciertos rasgos de las pruebas de inteligencia pueden contribuir a las calificaciones más bajas de diferentes nacionalidades y culturas. Por ejemplo, las sociedades analfabetas no siempre comparten el énfasis de las sociedades occidentales en cuanto a la velocidad, el resolver un problema con el menor número de pasos, la superioridad de las manipulaciones mentales en comparación con las físicas, o que la originalidad es mejor que la conformidad (Gill y Keats, 1980). A diferencia de la orientación más centrada en el tiempo y en sí mismas de las culturas occidentales, es más probable que las personas de sociedades muy tradicionales asocien la inteligencia con la gradualidad y la paciencia y que enfaticen la cooperación, la sociabilidad y el honor (Wober, 1974).

Entre otras diferencias culturales que pueden tener cierto efecto sobre las calificaciones de las pruebas se encuentra la perspectiva confuciana de la cultura china tradicional, la cual ve a la inteligencia como benevolencia y hacer lo correcto, y la perspectiva taoísta de la inteligencia que incluye la humildad, la libertad de estándares convencionales de juicio, y el conocimiento de uno mismo y de las condiciones externas (Yang y Sternberg, 1997). Los materiales de las pruebas de inteligencia también pueden ser percibidos de manera diferente por culturas distintas. Por ejemplo, Ortar (1963) encontró que cuando se les mostraba una ilustración de una cabe-

za sin boca los niños inmigrantes orientales en Israel tenían mayor probabilidad que los niños nativos de Israel de decir que faltaba el cuerpo. Y cuando se pidió a la gente de las tierras altas de Nueva Guinea que usaran un conjunto de cubos para copiar un diseño de dos dimensiones, muchos intentaron usar tanto la parte superior como los lados de los cubos.

Raza y grupo étnico. Uno de los temas más controvertidos en la medición de las habilidades cognoscitivas atañe a las diferencias raciales en el CI. Un hallazgo general de la investigación en este tema es que, aunque por lo regular se ha encontrado que el CI de los asiáticoamericanos es igual o mayor que el de los caucásicos, los CI promedio de los nativos americanos, los hispanoamericanos y los afroamericanos son significativamente menores. Entre las varias comparaciones de grupo, la atención se ha concentrado en las diferencias entre blancos y negros, una cuestión que se relaciona con la controversia herencia-ambiente.

Diferencias entre negros y blancos. Muchos científicos sociales (Klineberg, 1963; Lee, 1951) han atribuido los resultados de la investigación sobre las diferencias raciales en las habilidades cognoscitivas a las diferencias en los ambientes culturales de los niños negros y blancos; otros creen que las diferencias tienen una base genética (Eysenck, 1971; Jensen, 1969). Después de analizar los hallazgos de la investigación sobre las diferencias entre negros y blancos en la inteligencia, Jensen (1969) concluyó que la frecuencia de los genes que portan mayor inteligencia es menor en la población negra como un todo que en la blanca. La consecuencia, sostenía, era que los negros, aunque iguales a los blancos en la habilidad para la memorización, son más pobres en el razonamiento abstracto y la resolución de problemas.

Un conjunto de hallazgos empíricos citados por Jensen (1981) para refutar una explicación ambientalista estricta de las diferencias raciales en la inteligencia es que los niños hispanoamericanos e indios americanos que viven en condiciones ambientales aún peores que los negros tienen calificaciones promedio más altas en las pruebas de inteligencia no verbal. Además, a pesar de que sus padres y abuelos fueron sometidos a una severa discriminación en los siglos XIX y XX, las personas de origen chino y japonés en Estados Unidos superaban a los caucásicos en las calificaciones promedio de las pruebas no verbales de inteligencia, así como en los logros educativos y ocupacionales, y los igualaban en las calificaciones en pruebas de inteligencia verbal. Por último, los judíos, para quienes la discriminación social no es desconocida, de manera consistente han calificado más alto que otros grupos en medidas de inteligencia verbal (Vernon, 1985). Sin embargo, en muchos de esos grupos las tradiciones culturales y las características familiares alientan el alto rendimiento incluso cuando el legado nativo no sea necesariamente superior.

A pesar de los argumentos de Jensen (1980, 1981), Herrnstein y Murray (1994) y otros, la cuestión de las diferencias raciales en la inteligencia está lejos de ser resuelta. Los hallazgos de la investigación indican que los blancos superan a los negros en alrededor de una desviación estándar tanto en la WAIS-R (Reynolds *et al*., 1987) como en la Stanford-Binet: cuarta edición (Thorndike, Hagen y Sattler, 1986). Sin embargo, existe un traslape considerable entre las distribuciones de CI de los dos grupos étnicos: se estima que 15% de los negros obtiene CI más altos que los de los blancos promedio, y 15% de los blancos califica más bajo que la persona negra promedio (Vernon, 1985). Esas diferencias raciales en las calificaciones en las pruebas de inteligencia son atribuibles a una combinación interactiva de factores, incluyendo las deficiencias de las pruebas, diferencias en los entornos y diferencias genéticas, pero no se ha determinado la importancia relativa de cada una de esas tres fuentes de variabilidad.

Es de notar que la diferencia promedio entre las calificaciones de los blancos y los negros en las pruebas de inteligencia y aprovechamiento académico disminuyó casi la mitad de 1970 a

1990. Las explicaciones posibles para el estrechamiento de la brecha racial son los incrementos en el gasto en educación y la mayor educación de los padres, sobre todo entre los negros en los años recientes (Williams y Ceci, 1997).

Diferencias entre japoneses y estadounidenses. También relevante para la cuestión de las diferencias de nacionalidad y grupo étnico en la inteligencia es el hallazgo de CI promedio más altos en los niños japoneses que en los estadounidenses (Lynn, 1982). Durante muchos años se ha sabido que los hijos de inmigrantes asiáticos a Estados Unidos tienden a calificar al menos tan alto como los niños caucásicos en este país. Lynn (1982) informó que la diferencia en el CI promedio entre estadounidenses y japoneses criados en sus propios países era de alrededor de 11 puntos a favor del último grupo. De hecho, se ha estimado que al menos 10% de la población japonesa, en comparación con sólo 2% de los estadounidenses y europeos, tiene CI de 130 o mayores.

Se han ofrecido varias explicaciones posibles para tratar de comprender la diferencia en los CI promedio de niños japoneses y estadounidenses, una diferencia que se ha informado aumenta de manera gradual desde la Segunda Guerra Mundial. Suponiendo que las muestras de niños japoneses y estadounidenses a los que se examinó fueran igualmente representativas de las poblaciones específicas y que las pruebas fueran apropiadas por igual, la explicación más obvia tiene que ver con las diferencias entre las dos culturas en cuanto a las prácticas de crianza y educación formal de los niños. Una explicación biológica del aumento en el CI entre los japoneses es que, debido a las mejoras en salud y nutrición, los niños japoneses de la actualidad están mejor física y mentalmente que sus contrapartes en los días previos a la Segunda Guerra Mundial. Otra sugerencia es que los incrementos en el CI han sido causados por la heterosis (vigor híbrido) resultante de cierta disminución en los matrimonios consanguíneos (de parentesco) a medida que después de la Segunda Guerra Mundial grandes cantidades de japoneses se mudaron de pequeñas aldeas a grandes ciudades. Por último, Lynn (1987) propuso que las diferencias en inteligencia entre los caucásicos y las personas con antecedentes asiáticos se deben a diferencias genéticas en el funcionamiento del cerebro. Sostenía que en las personas de antecedentes asiáticos el hemisferio cerebral izquierdo evolucionó a estructuras capaces de procesar información visoespacial. El resultado de esta evolución, de acuerdo con Lynn, es que en los asiáticos una proporción mayor del tejido cortical se dedica al procesamiento de la información espacial y una proporción más pequeña está disponible para la información verbal. En consecuencia, la comunicación lingüística, como en la lectura y escritura de kanji, involucra habilidades espaciales que de manera normal dependen del hemisferio cerebral derecho. Por muy razonable que pueda parecer esta explicación de las mayores calificaciones obtenidas en las pruebas por los niños japoneses, Brody (1992) concluyó que la evidencia a favor de la teoría de Lynn no es convincente.

FACTORES BIOLÓGICOS Y HABILIDADES MENTALES

Los científicos modernos reconocen que el cerebro es el órgano de la actividad mental, pero los esfuerzos por identificar estructuras o áreas cerebrales específicas que son responsables de las habilidades cognoscitivas no han tenido mucho éxito. Con respecto al tamaño global del cerebro, algunos de los cerebros más pequeños de los que se tiene registro han sido de genios reconocidos (por ejemplo, Walt Whitman y Anatole France), y algunos de los cerebros más grandes han pertenecido a individuos con retraso severo. Aun así, varias revisiones de investigaciones han concluido que el tamaño global del cerebro tiene una pequeña correlación positiva con la habilidad intelectual (Broman, Nichols, Shaughnessy y Kennedy, 1987; Jensen y Sinha, 1991; Stott, 1983; Willerman, Schultz, Rutledge y Bigler, 1989). En un estudio de 139 infantes que tuvieron

bajo peso al nacer (menos de 1.5 kilogramos) se encontró que la circunferencia de la cabeza era un predictor importante del CI en la escala Stanford-Binet a los tres años de edad (Hack y Breslau, 1985). Esto siguió siendo cierto aun cuando se controlaron de manera estadística variables médicas y sociodemográficas, las cuales tenían relaciones significativas pero menores que la circunferencia de la cabeza con el CI posterior. Aunque el crecimiento compensatorio del cerebro durante los primeros ocho meses después del nacimiento compensó la disminución de los CI posteriores en algunos infantes, después de los ocho meses se observó poco crecimiento del cerebro. De este modo parecería que, al menos en los infantes, el tamaño de la cabeza puede anticipar la condición intelectual posterior (vea Wilson, 1985).

Localización cerebral de las funciones cognoscitivas

Podríamos desear que fuera posible hacer mejoras significativas en la inteligencia empleando técnicas quirúrgicas o químicas, pero en el presente eso es sólo ciencia ficción. Una hipótesis popular, que los procesos mentales de orden superior tienen lugar en los lóbulos frontales del cerebro, ha recibido cierto apoyo de los datos de los exámenes PET (tomografía por emisión de positrones) (Haier, 1991). El hallazgo temprano de que los pacientes sometidos a lobotomías prefrontales mostraban cierto deterioro postoperatorio en habilidades intelectuales específicas es congruente con dicha hipótesis (DeMille, 1962).

Los cambios en habilidades cognoscitivos específicas también están asociados con lesiones en otras áreas del cerebro. Por ejemplo, el daño del lóbulo temporal izquierdo —el hemisferio dominante en la mayoría de la gente— deteriora el desempeño verbal-simbólico más que el perceptual-espacial. Sin embargo, el daño del lóbulo temporal derecho afecta el desempeño perceptual-espacial más que el verbal-simbólico. Al evaluar los efectos del daño cerebral también debe considerarse la edad del paciente. El desarrollo intelectual de un niño pequeño puede resultar mucho más afectado por el mismo tipo de lesión cerebral que no tiene efecto mensurable en las habilidades intelectuales de una persona mayor.

Diferencias sexuales

En ocasiones se encuentran diferencias entre las calificaciones promedio de las pruebas de inteligencia de hombres y mujeres, pero por lo regular son intrascendentes. Sin embargo, los resultados de la investigación indican que hay diferencias sexuales en habilidades cognoscitivas y perceptual-motrices específicas. Halpern (1997) concluyó que a las mujeres les va mejor que a los hombres en tareas que requieren acceso y uso rápido de información fonológica, semántica y de otro tipo en la memoria a largo plazo. También destacan en tareas que requieren destreza motriz fina, velocidad perceptual y decodificación de información no verbal; tienen mejor articulación del habla y menores umbrales perceptuales para el tacto, el sabor y el olor. Por otro lado, los hombres se desempeñan mejor que las mujeres en tareas que involucran el razonamiento fluido, transformaciones en la memoria de trabajo visual o mover objetos, y en tareas motrices que requieren puntería. En lo que respecta a lo académico, las mujeres obtienen mayores calificaciones en la escuela, en particular en literatura y lenguas extranjeras. Los hombres se desempeñan mejor que las mujeres en pruebas de conocimiento en general y en geografía, matemáticas y ciencia. Esos hallazgos son, al menos en parte, función de las diferencias en la forma que nuestra sociedad trata a los niños y a las niñas. Por ejemplo, por lo regular se espera que las niñas tengan más logros en habilidades sociales y lingüísticas, mientras se supone que los niños deben desempeñarse mejor en matemáticas, mecánica y tareas con problemas relacionados.

Se ha encontrado que no sólo el sexo (género) sino también las hormonas sexuales están relacionadas con las habilidades cognoscitivas. Por ejemplo, Hier y Crowley (1982) encontraron una correlación positiva entre la habilidad espacial y las hormonas sexuales masculinas durante la pubertad. Los hallazgos de la investigación también sugieren que la testosterona vuelve más lento el desarrollo del hemisferio izquierdo y facilita el desarrollo del hemisferio derecho del cerebro, el cual está asociado con los tipos de habilidades de razonamiento que se necesitan para resolver problemas matemáticos (Christiansen y Knussmann, 1987). También es de interés el hallazgo de que las mujeres tienen un mejor desempeño en las pruebas de coordinación motriz y destreza verbal, pero un desempeño más pobre en las pruebas de razonamiento espacial, durante los momentos del mes en que los niveles de estrógeno en la sangre se encuentran en su punto máximo (Hampson, 1990; Kimura y Hampson, 1993). Las calificaciones de los hombres en las habilidades espaciales también fluctúan con sus niveles de testosterona: son más altas en la mañana que en el transcurso del día, y más altas en otoño que en primavera (Kimura y Hampson, 1994; Moffat y Hampson, 1996).

Se han ofrecido varias explicaciones neuropsicológicas para las diferencias sexuales en habilidades cognoscitivas específicas. Un conjunto de tales explicaciones apunta hacia el dimorfismo sexual en las estructuras nerviosas del hipotálamo, la amígdala y la corteza cerebral. Las mujeres tienen áreas de lenguaje que en proporción son más grandes que las de los hombres (Harasty, Double, Halliday, Kril y McRitchie, 1997), y se reporta que la densidad de las neuronas en las áreas de lenguaje de las mujeres es mayor que en los hombres (Witelson, Glezer y Kigar, 1995). Los cerebros de las mujeres también están organizados de una manera más bilateral que en los hombres, ya que en las mujeres las funciones cognoscitivas son menos específicas a un hemisferio cerebral particular. Además, el cuerpo calloso es más grueso en las mujeres que en los hombres, lo que permite una mejor conductividad entre los dos hemisferios cerebrales (Innocenti, 1994; Jancke y Steinmetz, 1994; Johnson, Pinkston, Bigler y Blatter, 1996). Por último, los datos de exámenes de tomografía por emisión de positrones (PET) indican que las áreas del cerebro en las que tiene lugar la mayor actividad mientras el individuo realiza funciones cognoscitivas específicas son diferentes en las mujeres y los hombres (Shaywitz *et al*., 1995).

Dieta y sustancias químicas

Desnutrición. La suposición de que la desnutrición fetal e infantil tiene efectos persistentes en la inteligencia es apoyada por numerosas investigaciones (por ejemplo, Lucas, Morley, Cole, Lister y Leeson-Payne, 1992; Zeskind y Ramey, 1981). Los intentos por revertir los déficit en la inteligencia relacionados con la desnutrición complementando las dietas de los niños desnutridos y exponiéndolos a un ambiente que les ofrezca cuidados no han tenido éxito del todo, aunque dicha intervención puede ayudar a detener esos déficit (Barba, 1981; Zeskind y Ramey, 1981).

Trastornos genéticos y dieta. La inteligencia muy baja se encuentra en individuos que padecen ciertos trastornos genéticos raros que son afectados por la dieta. En la fenilcetonuria (PKU), un trastorno genético causado por la falta de un gen que dirige la producción de una enzima responsable de oxidar la fenilalanina, la fenilalanina se acumula en la sangre y da lugar a una disminución drástica de las habilidades intelectuales. La PKU puede detectarse al momento del nacimiento con una prueba médica sencilla y, en consecuencia, el deterioro de la inteligencia puede ser prevenido cuando se coloca al niño en una dieta libre de fenilalanina.

La PKU y otros trastornos genéticos caracterizados por baja inteligencia, por ejemplo la enfermedad de Tay-Sachs y la galactosemia, se transmiten por genes recesivos. La enfermedad

de Tay-Sachs se asocia con una acumulación de una sustancia grasosa en el sistema nervioso central, mientras que la galactosemia se asocia con una acumulación de galactosa en la sangre. Al igual que la PKU, la galactosemia puede tratarse colocando al paciente en una dieta especial libre de galactosa.

Alcohol. Existen muchos *teratógenos* diferentes, drogas que pueden cruzar la barrera placentaria en una mujer embarazada y afectar el crecimiento y funcionamiento del cerebro del feto. El alcohol es una de esas drogas que, incluso cuando es consumido por una mujer embarazada en cantidades relativamente moderadas, puede contribuir a generar problemas de atención y tiempo de respuesta en los niños pequeños. Los efectos de la exposición prenatal a grandes cantidades de alcohol son todavía más graves, y dan por resultado una condición conocida como *síndrome fetal de alcohol* (*SFA*). Además del retraso en el crecimiento, apariencia facial distorsionada y malformaciones del cerebro y el cráneo, se presenta retraso mental en un gran porcentaje de los casos de SFA. De hecho, una de las causas más importantes de retraso mental en el mundo occidental es la exposición prenatal al alcohol. Por esta razón, se considera aconsejable que las mujeres embarazadas se abstengan por completo de beber alcohol (vea Spohr y Steinhausen, 1996; Streissguth, Bookstein y Barr, 1996).

Plomo. Otra sustancia que se ha demostrado tiene un efecto deteriorante en la inteligencia de los niños pequeños es el plomo, el cual existe en las viviendas, la comida, la tierra y el aire (Needleman, Gunnoe, Leviton y Perie, 1978; Needleman, Schell, Bellinger, Leviton y Allred, 1990; Thatcher, Lester, McAlaster, Horst e Ignasias, 1983). Needleman *et al.* (1990) demostraron la persistencia del defecto mental relacionado con el plomo en la adultez al reexaminar a 132 de 270 jóvenes adultos que habían sido examinados inicialmente cuando estaban en la escuela primaria. Se encontró que los individuos con mayores niveles de plomo con más frecuencia no habían logrado graduarse de secundaria y presentaban un ausentismo elevado; también tenían una incidencia más alta de problemas con la lectura y bajas calificaciones en las pruebas que miden vocabulario, razonamiento gramatical, habilidades motrices finas y coordinación ojo-mano. Esos hallazgos, combinados con los de otros investigadores (por ejemplo, Fulton *et al*., 1987; McMichael *et al*., 1988) apoyan la hipótesis de que la exposición a niveles elevados de plomo durante la niñez temprana tiene un efecto adverso sobre el desarrollo intelectual. La buena noticia es que los niveles de plomo en sangre de niños de uno a cinco años disminuyeron de manera considerable en las dos o tres décadas pasadas, una disminución atribuible en gran medida a la legislación que prohíbe el uso de plomo en las pinturas y tuberías y a la retirada progresiva del plomo en la gasolina (America's Children, 1998).

Herencia

La creencia en la determinación genética de la inteligencia se remonta al menos hasta la época de Francis Galton a finales del siglo XIX. Alfred Binet no rechazaba la idea de que la inteligencia estuviera genéticamente determinada, pero estaba más interesado en la posibilidad de modificar las habilidades intelectuales por medio de la educación, el entrenamiento y la intervención ambiental (Eysenck, 1984). Uno de los defensores más francos de la noción de que la inteligencia es determinada en gran medida por la herencia fue el psicólogo H. H. Goddard, quien defendía la reconstrucción de la sociedad a lo largo de las líneas del CI (Goddard, 1920).

La mayoría de los psicólogos, especialistas en el desarrollo infantil e investigadores educativos, probablemente estarían de acuerdo en que la inteligencia general, o al menos una predisposición al desarrollo cognoscitivo, es hasta cierto punto heredada (Snyderman y Rothman,

1987). Algunos investigadores genetistas consideran a la inteligencia como una *característica poligénica*, es decir, que es determinada por la interacción de muchos genes menores en lugar de un solo gen importante.

Quizá el método menos ambiguo de obtener información concerniente a los efectos ambientales sobre las habilidades cognoscitivas sea el de conducir un experimento con pares de gemelos monocigóticos (idénticos), quienes tienen herencias idénticas. Algunos pares de gemelos serían separados al nacer asignándolos a ambientes diferentes, mientras que otros pares se mantendrían juntos en el mismo ambiente. El hallazgo de mayores diferencias en las habilidades medidas entre los pares de gemelos criados en ambientes diferentes que entre los criados en el mismo ambiente sería un apoyo para la hipótesis de que el ambiente influye en las habilidades cognoscitivas.

Debido a que la sociedad no permitiría que científicos incluso bien intencionados movieran a los niños como piezas de ajedrez, se han diseñado métodos no experimentales para evaluar los efectos relativos de la herencia y el ambiente. Un enfoque consiste en comparar, en diversas edades cronológicas, los CI de gemelos monocigóticos que han sido criados por separado. De esta manera, la herencia se mantiene efectivamente constante mientras que el ambiente varía, aunque de una manera asistemática y no controlada. Además, pueden compararse los CI de individuos que tienen diferentes herencias pero que viven en ambientes similares, como los hermanos no idénticos o niños no relacionados a los que se cría juntos. También pueden hacerse comparaciones entre los CI de personas que tienen diferentes relaciones hereditarias y a quienes se cría en ambientes diferentes, como los hermanos no idénticos e individuos no relacionados criados aparte.

A pesar de la dificultad para localizar pares de gemelos monocigóticos que hayan sido criados por separado, se dispone de resultados de una serie de investigaciones de este tipo (encontrará resúmenes en Bouchard, Lykken, McGue, Segal y Tellegen, 1990; Bouchard y McGue, 1981; Plomin y Foch, 1980). En general, se ha encontrado que las correlaciones entre los CI de gemelos monocigóticos criados juntos son casi siempre más altas que las de gemelos monocigóticos criados por separado. Por ejemplo, Bouchard *et al.* (1990) informaron de correlaciones entre los CI obtenidos en la Escala de Inteligencia para Adultos de Wechsler (WAIS) por gemelos monocigóticos de .88 para la escala verbal, .79 para la escala de desempeño y .88 para la escala completa; los valores correspondientes para los gemelos monocigóticos criados aparte fueron de .64, .71 y .69. Además, entre más cercana fuera la relación genética entre los individuos, más altas eran las correlaciones entre sus calificaciones en las pruebas de inteligencia. Bouchard y McGue (1981) mencionaron las correlaciones medianas entre los CI de personas con diferentes grados de parentesco que vivían juntas, siendo de .86 para gemelos monocigóticos, .60 para gemelos dicigóticos, .47 para hermanos, .42 para padres e hijos, .33 para cónyuges y .29 para hermanos adoptados/naturales. En lo que se supone es un reflejo de la influencia del ambiente en el CI, las correlaciones fueron más bajas para pares correspondientes de gemelos a los que se crió por separado.

Los genetistas poblacionales a menudo expresan los resultados de los estudios de las diferencias hereditarias en términos de un *índice de heredabilidad* (h^2), definido como la razón de la varianza de la calificación en la prueba debida a la herencia con la varianza de la calificación en la prueba debida a una combinación de herencia y ambiente. Aunque se ha informado de estimados de heredabilidad de hasta .72 (Plomin, 1990), los estimados promedio de h^2 para la inteligencia en la población general son de alrededor de .50. Esto significa que un estimado de 50% de la varianza en las calificaciones CI puede atribuirse a factores genéticos. Sin embargo, debe advertirse que esos números no dicen nada acerca de la importancia relativa de la herencia o el ambiente en la determinación de la inteligencia de un individuo específico; los coeficientes de heredabilidad sólo se aplican a las poblaciones.

Incluso el más ávido defensor de una base genética de la inteligencia por un lado, o el más acérrimo ambientalista por el otro, reconocen que *tanto* la herencia *como* el ambiente son importantes en la formación de las habilidades cognoscitivas. En este contexto el *ambiente* no sólo se refiere al ambiente psicosocial o de experiencia de la persona, sino también al ambiente biológico prenatal y posnatal (nutrición, accidentes y cosas similares). Una interpretación de los datos de investigación que tienen que ver con esta materia es que la herencia establece una especie de límite superior a la inteligencia, un límite que sólo puede alcanzarse en las condiciones ambientales óptimas (Weinberg, 1989). Un corolario de esta proposición es que entre más alto sea el límite superior determinado por la herencia para la inteligencia de una persona, mayores serán los efectos potenciales del ambiente.

Otra manera de evaluar los efectos diferenciales de la herencia y el ambiente en las habilidades cognoscitivas está representada por la investigación de la adopción, como puede apreciarse en los Estudios de Adopción de Minnesota (Scarr y Weinberg, 1983) y el Proyecto de Adopción de Texas (Horn, 1983). En esas investigaciones se compararon los CI de grandes muestras de niños adoptados con los de sus hermanos no adoptados y los de sus padres adoptivos y biológicos. Los hallazgos de Horn (1983) son típicos en que los CI de los niños adoptados (de tres a diez años de edad) a los que estudió estaban mucho más cercanos a los de sus madres biológicas, de quienes habían sido separados casi desde el nacimiento, que de los CI de sus padres adoptivos. Los CI de los adolescentes en el estudio de Scarr y Weinberg (1983) también mostraron una correlación más alta con los CI de sus madres biológicas que con los de sus madres adoptivas.

Otro hallazgo interesante es que los efectos de la herencia sobre la inteligencia tienden a aumentar con la edad, mientras que los efectos del ambiente, y en particular del ambiente compartido, tienden a disminuir con la edad (McGue, Bouchard, Iacono y Lykken, 1993). Un factor que contribuye a ello es que, a medida que los niños y los adultos envejecen, la parte del ambiente que tuvo más influencia al principio de la vida es reemplazada por otras experiencias no compartidas en la escuela, en las interacciones sociales con los compañeros, en el trabajo y en otras situaciones.

El hecho de que las influencias genéticas se vuelven incluso más significativas con la edad fue subrayado por los resultados del Estudio de Gemelos de Louisville (Wilson, 1983). En esta investigación de 500 pares de gemelos, los CI de gemelos monocigóticos se hicieron más similares, pero los de gemelos dicigóticos se hicieron menos similares, de la infancia a la adolescencia. Los resultados de los Estudios de Adopción de Minnesota (Scarr y Weinberg, 1983) son congruentes con los del Estudio de Gemelos de Louisville en el descubrimiento de que el ambiente familiar tiene cierto impacto en el CI, en particular durante la niñez temprana, pero que los efectos del ambiente familiar son sustancialmente menores que los de la herencia. Otro hallazgo, aquel de un coeficiente estimado de heredabilidad de .80 para las calificaciones en pruebas de inteligencia en una muestra de adultos con una edad promedio de 66 años (Pedersen, Plomin, Nesselroade y McClearn, 1992), indica que la herencia continúa ejerciendo una influencia profunda en las calificaciones CI obtenidas tarde en la vida.

RESUMEN

A los individuos con calificaciones en los extremos bajo y alto de la distribución de inteligencia se les conoce, respectivamente, como retrasados mentales o superdotados. Tanto las calificaciones en las pruebas de inteligencia como la conducta adaptativa son importantes en el diagnóstico del retraso mental. El retraso mental se clasifica, de acuerdo con su gravedad, en tres o cuatro

categorías. Tanto la genética como la experiencia son factores determinantes en el retraso mental, pero en la mayoría de los casos se desconoce la causa exacta.

El estereotipo tradicional de que los niños superdotados son físicamente débiles, poco sanos, con probabilidades de consumirse pronto e inestables en lo emocional es incorrecto para la mayoría de esos niños, sobre todo para los que son moderadamente superdotados. Sin embargo, se ha informado que los niños extremadamente superdotados presentan mayor probabilidad que el promedio de tener problemas sociales y emocionales. La aceleración, el uso de mentores, el enriquecimiento, las clases especiales y las escuelas especiales se encuentran entre los procedimientos empleados en la educación de los niños superdotados.

El desempeño creativo no es sólo una función de una inteligencia relativamente alta, sino también de la elevada motivación, el entrenamiento especial y quizá de otras capacidades psicológicas. Un problema importante en el desarrollo de medidas útiles de la creatividad es la definición de criterios adecuados para inducir el desempeño creativo. Las baterías de pruebas como las Pruebas de la Estructura del Intelecto de Guilford y las Pruebas Torrance de Pensamiento Creativo son ejemplos notables de instrumentos diseñados para evaluar la creatividad. Los resultados de la investigación reciente sugieren que ciertas clases de desempeño creativo están asociadas con trastornos del estado de ánimo, como la psicosis maniaco-depresiva.

Dado un ambiente familiar relativamente estable, nutrición adecuada y experiencias educativas apropiadas, las calificaciones de CI permanecen bastante estables después de la niñez temprana. Los resultados de estudios transversales describen que la inteligencia aumenta en la juventud y luego declina de manera gradual en la vejez; los estudios longitudinales encuentran menos declinación con la edad. La tasa de deterioro, o incluso de aumento en algunos casos, es una función de los tipos de actividades a los que se dedica la gente a lo largo de su vida: quienes continúan comprometidos en actividades intelectuales muestran menor declinación intelectual que quienes manifiestan menos interés en el aprendizaje continuo. La cuestión de si la inteligencia disminuye de manera abrupta en las últimas semanas o meses antes de la muerte en la vejez, la *caída terminal*, no se ha resuelto de manera concluyente.

Un tamaño grande de la familia se asocia con menores CI promedio, y los primogénitos tienden a ser superiores en lo intelectual a los que nacen después. El estatus ocupacional y la posición socioeconómica tienen una correlación positiva entre sí y con la inteligencia, pero no queda claro si las ventajas de pertenecer a una clase social más alta den por resultado niños con CI más elevados o si los CI más altos y la posición social elevada son consecuencias de factores genéticos. Otras variables demográficas asociadas con las calificaciones CI son la residencia urbana contra la rural, el nivel educativo, la nacionalidad y el grupo étnico. En lo que respecta a la educación, las actitudes o expectativas de los maestros concernientes a qué niños son capaces de tener logros también pueden jugar cierto papel en si los niños alcanzan su potencial.

No se ha encontrado un área específica del cerebro que se considere el asiento de la inteligencia. Sin embargo, la investigación sobre la localización cerebral de las funciones cognoscitivas ha encontrado que ciertas estructuras desempeñan papeles importantes en los procesos mentales de orden superior.

Los estudios no han revelado diferencias de género consistentes en la habilidad mental general, aunque cada sexo tiende a ser superior al otro en ciertas habilidades específicas. Las niñas son mejores en memorización, tareas lingüísticas, velocidad perceptual y precisión y cálculos numéricos. Los varones destacan en razonamiento matemático, capacidad visoespacial, habilidad mecánica y velocidad y coordinación de los movimientos corporales grandes. Las bases fisiológicas de esas diferencias no se entienden bien, pero parecen estar relacionadas con diferencias en el desarrollo y funcionamiento de los hemisferios izquierdo y derecho del cerebro. Las dife-

rencias en otras estructuras cerebrales y en el nivel de testosterona también parecen estar relacionadas con las diferencias de género en las habilidades cognoscitivas.

Se ha encontrado que varias hormonas y drogas están relacionadas con las habilidades mentales. En particular, llaman la atención los estudios del síndrome fetal de alcohol y los efectos de los altos niveles de plomo en la inteligencia de los niños. La desnutrición, en especial durante el último periodo prenatal o el periodo posnatal temprano, puede producir un menor CI. Además, ciertos trastornos con base genética (por ejemplo, PKU, enfermedad de Tay-Sachs y la galactosemia) asociados con bajos CI pueden ser tratados con dietas especiales si se detectan con la oportunidad suficiente.

Entre los varios problemas y controversias que rodearon a las pruebas de inteligencia durante buena parte del siglo XX, la cuestión más debatida ha sido la de las contribuciones relativas de la herencia y el ambiente al moldeamiento de las habilidades cognoscitivas. La evidencia de docenas de investigaciones destaca la relación de la herencia con la habilidad mental general, aunque no niega que la herencia y el ambiente son importantes e interactivos en sus efectos sobre la conducta inteligente. Este tema ha resultado particularmente controvertido por su asociación con la problemática de las diferencias raciales en la inteligencia.

Aunque los hallazgos de numerosas investigaciones han llevado a concluir que en una población con apareamiento clasificado el coeficiente de heredabilidad (la proporción de varianza en las calificaciones de las pruebas de inteligencia de la población general explicada por la herencia) es hasta de .70, también está claro que los ambientes biológico y psicosocial tienen influencias importantes en la inteligencia.

PREGUNTAS Y ACTIVIDADES

1. Describa los sistemas de clasificación para el retraso mental propuestos por la Asociación Estadounidense del Retraso Mental, La Asociación Nacional para los Niños Retrasados y la Asociación Psiquiátrica Estadounidense.

2. Dado que en Estados Unidos el método para diagnosticar el retraso mental, incluyendo el CI límite, varía de un estado a otro, ¿es posible que un niño sea retrasado mental en un estado y "limítrofe" o de "bajo promedio" en otro? ¿Qué consecuencias podría tener esto?

3. Se invierten más fondos del gobierno en la educación de los retrasados mentales que en la de los superdotados. ¿Está esto justificado? ¿Por qué sí o por qué no?

4. Para "probar" su habilidad creativa, trate de resolver los siguientes ejercicios:
   a. ¿Cuántos usos puede imaginar para un clip, una pelota de goma, un ladrillo, una percha de alambre, una regla de un pie de longitud o un mondadientes?
   b. Trate de imaginar cómo cambiarían las cosas si:
   Todos tuvieran tres brazos.
   Todos tuvieran seis dedos y no tuvieran pulgar en cada mano.
   Lloviera de manera constante durante seis meses al año y no lloviera los seis meses restantes.

   Compare sus respuestas con las de sus amigos y condiscípulos.

5. ¿Qué variables demográficas están relacionadas con las calificaciones obtenidas en las pruebas de inteligencia? ¿Cuáles de esas variables parecen ser más importantes? ¿Cuáles tienen una relación causal con la inteligencia?

6. ¿Qué factores biológicos se ha demostrado que afectan la inteligencia? ¿Cuáles de esos factores son los más importantes?

7. Diseñe un estudio para probar la hipótesis de que la diferencia entre los CI promedio de negros y blancos no es significativa. No se preocupe demasiado con la posibilidad real de efectuar su estudio, pero asegúrese de controlar las variables extrañas (de confusión).

8. En un resumen de las correlaciones promedio entre los CI de personas que tienen diferentes grados de parentesco, Bouchard y McGue (1981) mencionaron que la correlación mediana entre los CI de gemelos fraternos del mismo sexo criados juntos era de .60, y que la correlación mediana entre los CI de gemelos idénticos criados juntos era de .86. Una fórmula sugerida para calcular el índice de heredabilidad es:

$$h^2 = \frac{r_i - r_f}{l - r_f},$$

donde r_i es la correlación entre los CI de gemelos idénticos (monocigóticos), y r_f es la correlación entre los CI de gemelos fraternos (dicigóticos) del mismo sexo criados juntos. Utilice esta fórmula para calcular h^2 e interprete el resultado.

CAPÍTULO NUEVE

EVALUACIÓN DEL DESARROLLO Y NEUROPSICOLÓGICA

Durante casi 100 años, las pruebas de inteligencia se han usado con el propósito de identificar las habilidades que niños y adultos poseen para entender y realizar tareas educacionales y ocupacionales, entre otras. Estas pruebas resultaron bastante efectivas con niños de edad escolar, pero han probado ser menos útiles para evaluar las habilidades de infantes y preescolares. Además, las pruebas de inteligencia general no fueron diseñadas para medir más que habilidades motrices, sensorial-perceptuales, lingüísticas y otras habilidades específicas o para proporcionar otra cosa que índices crudos de habilidades cognoscitivas específicas como memoria, atención-concentración y pensamiento abstracto.

Las dificultades y demoras en el aprendizaje pueden deberse a una baja habilidad mental, a impedimentos sensoriales y motrices o a trastornos neurológicos de varios tipos. En consecuencia, además de las medidas de habilidad mental general, a menudo se aplican pruebas especiales de memoria, percepción, habilidades psicomotrices y otras habilidades para proporcionar una imagen diagnóstica más detallada de los individuos que no presentan un funcionamiento efectivo en la escuela, el trabajo o en otros lugares.

La mayoría de los instrumentos expuestos en este capítulo no se aplican tan a menudo como las pruebas estándar de inteligencia, pero proporcionan fuentes adicionales de información para entender a niños y adultos y planear programas y tratamientos especiales dirigidos a quienes experimentan dificultades para adaptarse a las demandas de la vida cotidiana.

Este capítulo y el siguiente se interesan en los instrumentos psicométricos que se aplican con frecuencia para obtener información más detallada sobre las habilidades humanas que la proporcionada por las pruebas de inteligencia general. Las pruebas descritas en este capítulo se usan más a menudo en contextos clínicos, educativos y de investigación, mientras que los instrumentos analizados en el capítulo 10 se aplican sobre todo en los contextos de negocios e industrias.

EVALUACIÓN DEL DESARROLLO DE INFANTES Y NIÑOS PEQUEÑOS

Los estudios sistemáticos del desarrollo humano, iniciados hacia finales del siglo XIX, fueron impulsados gracias a la preocupación expresada por escritores y reformadores sociales acerca del bienestar de los niños, en particular por lo concerniente a su salud y educación, y sobre todo

por la explotación a que eran sometidos en los lugares de trabajo y en otras partes. Esta preocupación dio lugar a un movimiento por el bienestar infantil y a una legislación y programas públicos dirigidos a proporcionar un trato más humano a los niños. Asociadas con el movimiento por el bienestar infantil estaban la nueva ciencia de la psicología del desarrollo y la investigación sobre las características físicas, cognoscitivas, emocionales y sociales de los niños. Para contribuir a esta investigación se diseñaron instrumentos y procedimientos con los cuales medir el desarrollo cognoscitivo, motriz, perceptual, emocional y social.

Problemas en la examinación de infantes y niños pequeños

Examinar a infantes (0 a 1 1/2 años) y a preescolares (1 1/2 a 5 años) puede ser difícil debido a que mantienen la atención por periodos cortos y tienen mayor susceptibilidad a la fatiga. Los niños pequeños también pueden carecer de la motivación necesaria para seguir las tareas de una prueba, las cuales con frecuencia evalúan características que son más bien inestables durante la niñez temprana. Por esas razones, la confiabilidad y la validez de las pruebas aplicadas a preescolares tienden a ser menores que las resultantes de pruebas diseñadas para escolares. Las pruebas de inteligencia infantil también tienden a presentar bajas correlaciones con las calificaciones obtenidas en pruebas de inteligencia aplicadas a los mismos niños años después, y no proporcionan una predicción muy precisa del desarrollo intelectual posterior.

Una razón de la baja correlación que se da entre las calificaciones en las pruebas de inteligencia infantil y las calificaciones en pruebas como la Escala de Inteligencia de Stanford-Binet aplicadas a una mayor edad estriba en las diferencias existentes en los tipos de tareas que se realizan en las dos clases de pruebas. Las pruebas de inteligencia infantil son, sobre todo, medidas del desarrollo sensoriomotriz, como la habilidad para levantar y voltear la cabeza, seguir con la mirada un objeto en movimiento y alcanzar o agarrar un objeto. En contraste, los reactivos de las pruebas de inteligencia del tipo Binet son de naturaleza más lingüística o verbal. Los niños preescolares, que tienen un repertorio conductual mayor que el de los infantes, pueden caminar y sentarse en una mesa mientras manipulan los materiales de la prueba, y se comunican mejor con el examinador.

Las pruebas de inteligencia infantil no sólo tienen una validez predictiva relativamente baja, sino que su confiabilidad también es menor que la de las pruebas aplicadas más tarde durante el periodo preescolar. Aunque la mayor tendencia a la distracción de los infantes en situación de prueba contribuye a la baja confiabilidad de los instrumentos que se les aplican, de buena fe se afirma que al parecer también ocurren cambios en las habilidades cognoscitivas de los niños pequeños.

Los niños no sólo se muestran más atentos y motivados que los infantes en las situaciones de prueba, sino que sus habilidades cognoscitivas parecen ser de una calidad diferente. Por ejemplo, los preescolares se interesan mucho más en las palabras y las interacciones sociales que los infantes.

A pesar de sus bajas correlaciones con los resultados de pruebas posteriores, las pruebas aplicadas durante la infancia son útiles para diagnosticar el retraso mental y los trastornos cerebrales orgánicos, y en la detección de las discapacidades del desarrollo. Los hallazgos de la investigación han revelado que las calificaciones obtenidas en las pruebas durante la infancia proporcionan una predicción significativa de la condición intelectual posterior de niños con retraso mental y con daño neurológico (Ames, 1967; McCall, 1979). Aunque los resultados de dichos estudios indican que el desempeño en las pruebas infantiles puede contribuir a la comprensión del desarrollo del niño y a tomar decisiones prácticas acerca de este grupo de edad, los datos de prueba deben combinarse e interpretarse a la luz de otra información acerca del examinado y teniendo conciencia de las limitaciones de las pruebas.

Programas de Desarrollo de Gesell

La investigación iniciada por Arnold Gesell en la Clínica Yale de Desarrollo Infantil durante la década de 1920 dio lugar a una serie exhaustiva de investigaciones sobre la infancia y la niñez temprana que continuaron durante 40 años. Una suposición que guiaba esos estudios era que las funciones motrices gruesas y finas, de lenguaje, personal-sociales y de conducta adaptativa de los niños seguían una secuencia ordenada de maduración. Se obtuvieron datos normativos sobre el desarrollo de las habilidades motrices, lingüísticas y personal-sociales, así como de la conducta adaptativa desde el nacimiento hasta los seis años. Se obtuvo información detallada de cada niño siguiendo diversos métodos: registros en el hogar, historia médica, registros diarios, mediciones antropométricas, observaciones materiales, informes del comportamiento del niño en la clínica, examinación normativa y calificaciones del desarrollo. El siguiente extracto es característico de las descripciones conductuales normativas proporcionadas por Gesell y sus colaboradores (Gesell y Amatruda, 1941, p. 41):

> El bebé puede alcanzar con sus ojos antes de poder alcanzar con su mano; a las 28 semanas un bebé mira un cubo; lo agarra, siente la superficie y el borde conforme lo empuña, lo lleva a su boca, donde siente sus cualidades de nuevo, lo aparta, lo mira al alejarlo, lo hace girar mientras mira, mira mientras lo hace girar, lo regresa a su boca, lo retira de nuevo para inspeccionarlo, lo regresa una vez más a la boca, lo cambia a la otra mano, lo golpea, lo toca con la mano libre, lo cambia, lo lleva de nuevo a la boca, lo deja caer, lo recupera, lo lleva otra vez a la boca, repitiendo el ciclo con variaciones —todo en el tiempo que se lleva leer esta frase.

Las calificaciones en los Programas de Desarrollo de Gesell, determinadas por la presencia o ausencia de conductas específicas características de los niños a ciertas edades, se resumieron en términos de la *edad de desarrollo* (*ED*). La ED podía ser convertida luego a un *cociente de desarrollo* (*CD*) mediante la fórmula *CD* = 100 (*ED/EC*). Sin embargo, Gesell no consideró que el CD fuera equivalente a un CI.

Es probable que los Programas de Desarrollo de Gesell fueran más usados por los pediatras que por los psicólogos de la década de 1920 hasta la de 1940. Los psicólogos, en particular los que tenían una orientación psicométrica fuerte, criticaban la subjetividad y la mala estandarización de los programas de Gesell. Sin embargo, una versión posterior de las escalas incluía procedimientos observacionales más objetivos. Knobloch (Knobloch y Pasamanick, 1974; Knobloch, Stevens y Malone, 1987) proporcionó instrucciones detalladas para efectuar observaciones e interpretarlas en la revisión de los Programas de Desarrollo de Gesell. También se publicaron normas para preescolares ($2^1/_2$ a 6 años) con intervalos de medio año, pero no para infantes (Ames, Gillespie, Haines e Ilg, 1979).

Los Programas de Desarrollo de Gesell fueron populares, sobre todo entre los pediatras, y todavía están en uso revisiones de los programas originales (Ireton, 1992, 1998). Sin embargo, los psicólogos del desarrollo perseveraron para elaborar instrumentos con mejores características psicométricas que las de los programas de Gesell. Algunos ejemplos son la Escala Mental de California para el Primer Año, la Prueba de Inteligencia Northwestern, la Escala Griffith del Desarrollo Mental, la Escala Merrill-Palmer y la Escala Cattell de Inteligencia Infantil. Sólo las dos últimas siguen imprimiéndose, y en su mayor parte el contenido ha sido reemplazado.

Un derivado más reciente de los Programas de Desarrollo de Gesell es el programa Denver-II, (de W. K. Frankenburg *et al*.; Denver Developmental Materials). El Denver-II fue diseñado para evaluar las habilidades personales, sociales, motrices finas y gruesas, de lenguaje y

adaptativas de los niños desde el nacimiento hasta los seis años, y funciona como instrumento de detección de las demoras del desarrollo. Los 125 reactivos del Denver-II se administran de manera individual en 20 a 25 minutos, o en 10 a 15 minutos en la versión abreviada. Se califica en cuatro áreas: personal-social, motriz fina-adaptativa, lenguaje y motriz gruesa. También se obtienen calificaciones en cinco conductas: típica, docilidad, interés en los alrededores, timidez y lapso de atención. El Denver-II es fácil de administrar y de calificar, pero se le ha criticado por la poca representatividad de su muestra de estandarización (Hughes, 1995).

Escala Brazelton de Evaluación Conductual Neonatal

A lo largo de su vida, la gente es evaluada de muchas maneras, formales e informales, y en ocasiones incluso antes de nacer. Por ejemplo, la Escala Obstétrica Rochester consta de una escala prenatal, una escala para el parto y una escala infantil. Otra medida, la calificación Apgar, se deriva de mediciones del ritmo cardiaco, la respiración, el tono muscular, los reflejos y el color obtenidas al minuto y a los cinco minutos del nacimiento (Chinn, Drew y Logan, 1975). Sin embargo, es posible que la prueba neonatal más popular sea la Escala Brazelton de Evaluación Conductual Neonatal (NBAS) (Brazelton, 1973, 1984).

La NBAS, que tiene un rango de edad de tres días a cuatro semanas, se califica en 26 reactivos conductuales y 20 respuestas provocadas, incluyendo medidas del funcionamiento neurológico, conductual y social. Los reactivos miden la coordinación mano-boca, la habituación a los estímulos sensoriales, las respuestas de sobresalto, reflejos, respuestas a la tensión, madurez motriz y caricias . A pesar de ciertos defectos, por ejemplo, pocos datos normativos o de validez y coeficientes de confiabilidad bastante bajos, la NBAS sigue siendo usada por los pediatras y los psicólogos infantiles en la práctica y la investigación.

Escalas de Bayley del Desarrollo Infantil

Las Escalas de Bayley del Desarrollo Infantil, segunda edición (BSID-II) (The Psychological Corporation), están basadas en el Estudio de Crecimiento de Berkeley, un programa de investigación dirigido por Nancy Bayley. La BSID-II fue diseñada para niños de entre uno y 42 meses de quienes se sospeche que están en riesgo de presentar discapacidades cognoscitivas y consta de tres partes: una Escala Mental que arroja un Índice de Desarrollo Mental, una Escala Motriz que produce un Índice de Desarrollo Psicomotriz y una Escala de Calificación de la Conducta que complementa la información de las escalas mental y motriz. La Escala Mental mide las habilidades sensorial-perceptuales, discriminaciones y la habilidad de responder a ellas; la adquisición de constancia del objeto; memoria, aprendizaje y resolución de problemas; vocalización, inicio de la comunicación verbal, evidencia temprana de la base del pensamiento abstracto, habituación, mapeo mental, lenguaje complejo y formación de conceptos matemáticos. La Escala Motriz mide el grado de control corporal, coordinación de los músculos grandes, habilidades manipulatorias finas de las manos y los dedos, movimiento dinámico, práctica dinámica, imitación postural y estereognosis. La Escala de Calificación de la Conducta mide atención-activación, orientación-compromiso, regulación emocional y calidad motriz. La prueba entera puede administrarse en 25 a 35 minutos a niños menores de 15 meses y en un máximo de 60 minutos a niños mayores de esa edad. (Vea la figura 9.1.)

La BSID-II fue estandarizada a principios de la década de 1990 en 850 niños y 850 niñas, de 1 a 42 meses de edad, seleccionados de manera aleatoria estratificada de cuatro regiones geo-

FIGURA 9.1 Reactivos para las Escalas de Bayley de Desarrollo Infantil, segunda edición.

(Copyright © 1993 por The Psychological Corporation, una Harcourt Assessment Company. Reproducido con autorización. Todos los derechos reservados. "Escalas de Bayley de Desarrollo Infantil" es una marca registrada de Psychological Corporation inscrita en Estados Unidos y otras jurisdicciones.)

gráficas y por edad, género, grupo étnico y educación de los padres. El manual de la BSID-II proporciona datos sobre niños que nacieron de manera prematura, en quienes la prueba de VIH resultó positiva, que fueron expuestos a drogas durante el periodo prenatal, que fueron asfixiados al nacer, que presentan demoras en el desarrollo o tienen infecciones frecuentes del oído medio, que son autistas o tienen síndrome de Down. Un instrumento acompañante, el Examen de Bayley de Neurodesarrollo Infantil (BINS), fue diseñado para evaluar las funciones neurológicas básicas, las funciones receptivas auditivas y visuales, y los procesos sociales y cognoscitivos en niños de 3 a 24 meses.

La MSCA y la MST

Las Escalas McCarthy de las Habilidades de los Niños (MSCA) (The Psychological Corporation), que comienzan donde terminan las escalas de Bayley, fueron diseñadas para niños de $2\frac{1}{2}$ a $8\frac{1}{2}$ años de edad. Estas escalas producen seis medidas de desarrollo intelectual y motriz: verbal, perceptual-desempeño, cuantitativo, cognoscitivo general, memoria y motriz. La MSCA fue estandarizada en muestras de alrededor de 100 niños en cada uno de diez grupos de edad, estratificados por raza, región, posición socioeconómica y residencia urbana-rural. Los datos sobre la validez de la MSCA, publicados después de la muerte de la autora, siguen siendo escasos.

La Prueba de Detección de McCarthy (MST), publicada años después de la MSCA, proporciona un medio para identificar a niños (de 4 a $8\frac{1}{2}$ años) que pueden estar en riesgo de presentar problemas de aprendizaje. Las seis escalas componentes de la MST se inspiraron en las de la MSCA.

FirstSTEP y el ESP

Las pruebas psicológicas usadas para detectar demoras en el desarrollo en grandes cantidades de niños y la subsecuente examinación diagnóstica a profundidad deberán cumplir con los criterios mencionados en el Acta para la Educación de Individuos con Discapacidades (IDEA) (Ley pública 101-476). Aunque la MSCA cumple los criterios de la IDEA, dos instrumentos diseñados específicamente con esas consideraciones en mente son la Prueba de Detección FirstSTEP para la Evaluación de Preescolares (The Psychological Corporation) y los Perfiles de Detección Temprana AGS (ESP) (American Guidance Service). Las características psicométricas de FirstSTEP y del ESP son aceptables para los instrumentos de detección del desarrollo, pero ningún instrumento ha sido usado de manera extensiva con fines de investigación.

FirstSTEP es una prueba rápida (15 minutos) para detectar demoras en el desarrollo en niños de 2.9 a 6.2 años de edad. Las 12 subpruebas, que fueron diseñadas para crear una atmósfera de "juego" en el examen, se clasifican en tres de los cinco dominios de la IDEA: cognición, comunicación y motriz. El desempeño del niño en las 12 subpruebas del FirstSTEP se expresa como una calificación compuesta interpretada en términos de tres categorías de clasificación; "dentro de límites aceptables", "precaución" (demoras en el desarrollo de leves a moderadas), o "en riesgo" (de sufrir demoras en el desarrollo). Las Escalas de Calificación Social-Emocional y Padres/Maestro son opcionales y se utilizan para evaluar el cuarto dominio de la IDEA (niveles de atención/actividad, interacciones sociales, rasgos personales, y problemas de conducta serios), y una Lista de Verificación de Conducta Adaptativa, también opcional, evalúa el quinto dominio de la IDEA (actividades de la vida cotidiana, autocontrol, relaciones e interacciones, y funcionamiento en la comunidad).

Los Perfiles de Detección Temprana AGS (ESP) son un inventario breve para determinar demoras en el desarrollo de los preescolares (de 2 años a 6 años 7 meses). Consta de tres componentes básicos (perfiles) y cuatro estudios complementarios. La aplicación de los perfiles se lleva menos de 30 minutos y los estudios necesitan de 15 a 20 minutos. El Perfil Cognoscitivo/Lenguaje consta de tareas para evaluar habilidades de razonamiento, organización visual y discriminación, vocabulario receptivo y expresivo, y destrezas escolares básicas. El Perfil Motriz evalúa habilidades motrices gruesas y finas (por ejemplo, caminar por una línea recta, imitar movimientos de brazo y pierna, trazar laberintos, dibujar formas). El Perfil de Autoayuda/Social, un cuestionario que es llenado por uno de los padres o por otro cuidador del niño, se interesa en el desempeño típico del niño en la comunicación, habilidades de la vida cotidiana, socialización y habilidades motrices. Los cuatro estudios del ESP son el Estudio de Articulación (el niño pronuncia 20 palabras), el Estudio del Hogar (los padres responden a preguntas acerca del ambiente familiar del niño), la Historia de Salud (los padres verifican los problemas de salud que ha tenido el niño) y el Estudio de Conducta (el examinador califica el lapso de atención, la tolerancia a la frustración, el estilo de respuesta y otras conductas del niño durante la aplicación de los perfiles Cognoscitivo-Lenguaje y Motriz). Las calificaciones en el ESP se convierten a índices de detección al nivel I o a calificaciones estándar, rangos percentilares y equivalentes de edad al nivel II, indicando si el niño requiere evaluación posterior.

Otras pruebas del desarrollo

Se dispone de otras baterías y pruebas específicas, nuevas o revisadas, para evaluar el desarrollo motriz, perceptual, cognoscitivo, emocional y social durante la infancia y la niñez temprana. Algunos de estos instrumentos son simples formas en las que un padre, tutor u otra persona familiarizada con el niño efectúa y registra observaciones de su conducta y sus características

cotidianas. Otros instrumentos implican la presentación de materiales al niño, a quien por lo general se le pide que haga algo con los materiales; las respuestas del niño se anotan y evalúan. Ciertos instrumentos psicométricos, como las Escalas de Desarrollo Motriz de Peabody, la Prueba del Desarrollo del Lenguaje-Primario, tercera edición, y la Prueba del Desarrollo de la Percepción Visual, segunda edición, todos los cuales pueden encontrarse en pro.ed, fueron diseñados para evaluar el desarrollo en dominios específicos. Otros instrumentos, como los que se describen a continuación, son baterías de pruebas para evaluar el desarrollo de un niño en varios dominios.

Evaluación del Desarrollo de Niños Pequeños (DAYC). La DAYC (pro.ed) identifica posibles demoras en el desarrollo cognoscitivo, comunicativo, social-emocional, físico y de conducta adaptativa durante los primeros seis años de vida. Esos cinco dominios reflejan áreas en las que el Acta para la Educación de Individuos con Discapacidades (IDEA) de 1990 ordena la evaluación e intervención. A cada uno de los cinco dominios corresponde una subprueba que, dependiendo de la edad del niño, puede aplicarse en 10 a 20 minutos. Las calificaciones de los cinco dominios proporcionan información sobre fortalezas y debilidades específicas, y distingue entre los niños que se desarrollan de manera normal y quienes presentan un desarrollo significativamente por debajo del normal. Las calificaciones también pueden usarse para documentar el progreso en las habilidades del desarrollo como resultado de programas específicos de intervención. Los datos de confiabilidad y validez para los cinco dominios de la DAYC y las calificaciones compuestas dadas en el manual (Voress y Maddox, 1998) son muy alentadoras con respecto a la DAYC como medida del desarrollo.

Evaluación del Desarrollo de Infantes y Niños Pequeños. La Evaluación del Desarrollo de Infantes y Niños Pequeños (IDA) (Riverside Publishing) es otro enfoque centrado en el dominio para la identificación de niños, desde el nacimiento hasta los 36 meses, que están en riesgo. Más que ser una batería de pruebas *per se*, IDA es un procedimiento comprensivo, multidisciplinario, centrado en la familia, que involucra a un equipo de profesionales para obtener, revisar e integrar datos de múltiples fuentes. El proceso de evaluación consta de seis fases, cada una de las cuales se desarrolla a partir de la precedente y es completada luego de ser analizada y revisada por el equipo. La fase 4 de los procedimientos de IDA, Fase de Observación y Evaluación del Desarrollo, hace uso del Perfil Provence de Desarrollo desde el Nacimiento hasta los Tres. La evaluación estandarizada del desarrollo proporcionada por el Perfil Provence emplea la observación naturalista e incorpora informes de los padres sobre el desarrollo del niño en ocho dominios: motriz gruesa, motriz fina, relación con objetos inanimados (cognoscitiva), lenguaje/comunicación, autoayuda, relación con personas, emociones y estados de ánimo (afectos), y afrontamiento. Los coeficientes de confiabilidad para las calificaciones en esos dominios fluctúan de la parte superior de .70 a la parte media de .90, dependiendo de la edad del niño. También se han presentado varios tipos de evidencia a favor de la validez de IDA (vea Erikson, 1995; Meisels y Fenichel, 1996).

DISCAPACIDADES DE APRENDIZAJE

Las dificultades para aprender a leer, escribir, deletrear o realizar operaciones aritméticas y otras habilidades académicas, de manera tradicional habían sido atribuidas a retraso mental, impedimentos físicos, problemas emocionales graves o falta de motivación. Pero incluso cuando se eliminan esas fuentes como posibles explicaciones, sigue existiendo un grupo considerable de niños que experimentan problemas en el aprovechamiento escolar. Se dice que esos niños tienen una discapacidad específica de aprendizaje o simplemente una *discapacidad de aprendizaje*

(*DA*). Las discapacidades de aprendizaje pueden ocurrir en individuos de cualquier nivel de inteligencia, pero, en contraste con el retraso mental, los logros de los niños con DA están significativamente por debajo de su capacidad cognoscitiva general.

Demografía y definiciones

Las discapacidades de aprendizaje constituyen la mayor condición de impedimento entre los niños de todo el mundo (Stanford y Oakland, 2000). A mediados de la década de 1990, un estimado de cinco millones o más de escolares y jóvenes estadounidenses tenían una o más discapacidades. A la mitad de esos estudiantes se les diagnosticó una discapacidad de aprendizaje. Entre los que tienen discapacidades de aprendizaje, los varones superan a las mujeres por dos a uno. Dentro de los grupos raciales-étnicos, el porcentaje de niños con discapacidades de aprendizaje es mayor para los indios americanos y más bajo para los asiáticos/isleños del Pacífico (U. S. Department of Education, 1997).

La estadounidense Ley Pública 101-476, Acta para la Educación de Individuos con Discapacidades (IDEA), de 1990, define las discapacidades de aprendizaje como:

> El término "niños con discapacidades específicas de aprendizaje" se refiere a aquellos niños que tienen un trastorno en uno o más de los procesos psicológicos básicos involucrados en la comprensión o en el uso del lenguaje, hablado o escrito, trastorno que puede manifestarse en una habilidad imperfecta para escuchar, pensar, hablar, leer, escribir, deletrear o para hacer cálculos matemáticos. Dichos trastornos incluyen condiciones como impedimentos perceptuales, lesión cerebral, disfunción cerebral mínima, dislexia y afasia del desarrollo. Dicho término no incluye a niños con problemas de aprendizaje que resultan sobre todo de impedimentos visuales, auditivos o motrices, de retraso mental, de perturbación emocional o de desventaja ambiental, cultural o económica.

El tipo más común de discapacidad para el aprendizaje es la *dislexia*, en la cual la persona tiene dificultades para leer en silencio o en voz alta. Cuando se le pide que lea en voz alta, un niño disléxico lo hace de manera lenta, vacilante y laboriosa. Los niños disléxicos experimentan dificultades en la lectura debido a problemas con la codificación fonológica (es decir, decodificar las letras impresas en sonidos mezclados). La dislexia, que es de tres a cuatro veces más común entre los varones que entre las mujeres, puede deberse a una incapacidad para procesar los sonidos (*dislexia auditiva*), a la dificultad para procesar la información que ha sido vista (*dislexia visual*) o a trastornos de comprensión o problemas con la producción escrita. Instrumentos como la Prueba de Detección de Dislexia y las Pruebas de Detección Temprana de Dislexia (de R. Nicholson y A. Fawcett; The Psychological Corporation) son útiles para identificar a escolares y preescolares disléxicos.

Los problemas de aprendizaje no verbal en matemáticas (*discalculia*), escritura (*disgrafia*) y cognición espacial son menos comunes que los problemas de aprendizaje verbal (Rourke, 1989). La dificultad en el aprendizaje de la aritmética puede estar relacionada con problemas de lenguaje o de lectura, así como con perturbaciones en el pensamiento cuantitativo, la visualización o escritura de números, y el recuerdo de instrucciones (Johnson y Myklebust, 1967). Sólo alrededor de 1 a 10% de las personas con discapacidades de aprendizaje presentan dichos problemas, en comparación con .1 a 1% de la población general.

Los niños con DA verbales, por lo general, tienen un mejor desempeño en las pruebas de ejecución, las cuales requieren destrezas visoespaciales y visomotrices, que en las pruebas verbales, las cuales miden las habilidades de lenguaje. Sucede lo opuesto en niños con DA no verbales: se desempeñan mejor en las pruebas verbales que en las de ejecución.

Causas de las discapacidades de aprendizaje

Existe un debate considerable acerca de si las DA son causadas por factores neurológicos, del desarrollo, de la experiencia o de una combinación de estos. Las condiciones neurológicas asociadas con las DA pueden atribuirse a influencias prenatales como los virus, el alcohol, a fumar cigarrillos o a drogas como la cocaína, a la radiación y a otros teratógenos que pueden cruzar la barrera placentaria y dañar al embrión o feto. El nacimiento prematuro, el bajo peso al nacer y el uso de fórceps también pueden participar en las discapacidades de aprendizaje (Bender, 1995). Los factores posnatales que han sido investigados como causas posibles de las DA son las convulsiones inducidas por fiebres altas o la inhalación de contaminantes con plomo (Needleman, Schell, Bellinger, Leviton y Allred, 1990); la diabetes, la meningitis, las lesiones en la cabeza y la desnutrición también han sido implicadas en ciertos casos (Hallahan, Kauffman y Lloyd, 1996).

Existe evidencia de una base genética para ciertas DA (por ejemplo, Oliver, Cole y Hollingsworth, 1991). Una línea relacionada de investigación neuropsicológica se ha centrado en déficit en el lóbulo temporal izquierdo del cerebro de la gente con discapacidades de aprendizaje verbal. Una estructura cerebral de interés es el plano temporal, un área en ambos lados del cerebro que se conoce por participar en el desarrollo del lenguaje. En los no disléxicos el plano temporal del lado izquierdo del cerebro es notablemente más grande que el del lado derecho, pero en los disléxicos no hay diferencia en el tamaño de los planos temporales en los dos lados del cerebro (Leonard *et al*., 1996).

Diagnóstico y tratamiento

En las aulas, los maestros pueden identificar las discapacidades de aprendizaje en los niños mediante la observación cuidadosa. También pueden aplicar pruebas colectivas de inteligencia y/o instrumentos más especializados como el Procedimiento de Calificación de la Discapacidad de Aprendizaje (Academic Therapy Publications), la Escala de Evaluación de Discapacidades de Aprendizaje (Hawthorne Educational Services), la Prueba de Detección de McCarthy y las Pruebas de Detección Slingerland para la Identificación de Niños con Discapacidad Específica de Lenguaje (Educators Publishing Service). Sin embargo, la administración de una batería de pruebas psicológicas requiere los servicios de un psicólogo escolar o un psicólogo clínico.

El diagnóstico efectivo y la planeación del remedio en las discapacidades de aprendizaje son una empresa multidisciplinaria que incluye al maestro regular del niño, a especialistas que tienen conocimientos relacionados con el impedimento sospechado y a personas experimentadas en el uso de instrumentos psicométricos para hacer evaluaciones diagnósticas. De acuerdo con las directrices proporcionadas por la Ley pública estadounidense 94-142, Acta de Educación para Todos los Niños con Impedimentos, de 1975, sólo se hace un diagnóstico de una discapacidad de aprendizaje específica cuando se encuentra una diferencia significativa entre la habilidad y el aprovechamiento en una o más de las siguientes áreas: expresión oral, comprensión auditiva, expresión escrita, habilidad básica de lectura, lectura de comprensión, cálculos matemáticos o razonamiento matemático.

Una vez que se cuenta con diagnóstico de una discapacidad de aprendizaje, debe prepararse un *plan de educación individualizada* (*PEI*) que consta de objetivos a corto y largo plazos y procedimientos para alcanzarlos. Además de un plan para remediar los déficit relacionados con la escuela, un PEI efectivo incluye medidas para tratar los problemas conductuales acompañantes.

En Estados Unidos, los criterios de elegibilidad para proporcionar servicios a los niños con discapacidades de aprendizaje varían de un estado a otro, pero, en general, el diagnóstico de una discapacidad de aprendizaje sólo queda justificado cuando la calificación global de un niño

en una prueba estandarizada de aprovechamiento está al menos una desviación estándar por debajo de su calificación en una prueba de inteligencia co-normada. Las pruebas individuales de inteligencia, como SB-IV, WPPSI-R, WISC-III y K-ABC, y las pruebas estandarizadas de aprovechamiento como la Prueba Peabody de Aprovechamiento Individual, revisada, el test Kaufman de Rendimiento Educativo y la Prueba Wechsler-II de Aprovechamiento Individual son apropiadas. Es factible que para este propósito se haya aplicado de manera más amplia la prueba Woodcock-Johnson III, la cual incluye una batería de pruebas de inteligencia (pruebas WJ-R de habilidad cognoscitiva) y una batería co-normada de pruebas de aprovechamiento (pruebas WJ-R de aprovechamiento). En los capítulos 6 y 7 se proporcionan descripciones de esas pruebas. Además de las baterías de pruebas de inteligencia y aprovechamiento, en ciertos casos es conveniente aplicar pruebas más especializadas de desarrollo neuropsicológico, mental e incluso pruebas de personalidad.

Se ha utilizado una variedad de procedimientos de instrucción en los niños con DA, incluyendo el análisis conductual e intervención, el aprendizaje cooperativo, la tutoría de pares y agresiva, y la asesoría en habilidades de razonamiento (Bender, 1995; Kirk, Gallagher y Anastasiow, 1997; Sullivan, Mastroipieri y Scruggs, 1995). Los resultados de esas y otras estrategias de intervención (por ejemplo, biorretroalimentación, entrenamiento de relajación, instrucción multisensorial, dietas especiales) han sido mixtos.

TRASTORNOS NEUROPSICOLÓGICOS Y EVALUACIÓN

En tiempos antiguos el oráculo de Delfos recomendaba a quienes buscaban su consejo que empezaran por conocerse a sí mismos, pero a pesar de la búsqueda e investigación continuas por casi dos siglos, esta tarea ha demostrado no ser sencilla. El funcionamiento del casi kilo y medio de tejido esponjoso que compone el cerebro humano en ocasiones parece ser casi tan complejo como el universo mismo. Con todo, ahora sabemos bastante acerca del funcionamiento de los cuatro lóbulos de la corteza cerebral (frontal, parietal, occipital, temporal) y las estructuras subcorticales del cerebro.

Si bien el pensamiento y la acción por lo regular involucran muchas áreas diferentes del cerebro, existe cierto grado de especificidad o localización en su funcionamiento. Por ejemplo, sabemos que en la mayoría de las personas un área del lóbulo frontal izquierdo (*área de Broca*) desempeña un papel importante en la producción del lenguaje gramatical, y que un área del lóbulo temporal izquierdo (*área de Wernicke*) le da significado al lenguaje. También sabemos que el lóbulo parietal izquierdo es importante en la orientación visoespacial, que los lóbulos frontales desempeñan un papel importante en el pensamiento abstracto y la resolución de problemas, y que el hipocampo participa en el almacenamiento de los recuerdos. Sin embargo, dependiendo de la edad del individuo y de otros factores, cuando un área particular del cerebro es lesionada, otras áreas pueden asumir el control de las funciones del área lesionada o compensar su pérdida.

Modelo Reitan-Wolfson

La figura 9.2 es un esbozo del marco de referencia conceptual del funcionamiento neuropsicológico propuesto por Reitan y Wolfson (1993) para la organización de los correlatos conductuales del funcionamiento cerebral y la descripción de medidas de esas funciones. El proceso comienza con la entrada de la información sensorial al cerebro. Esto es seguido por el primer paso

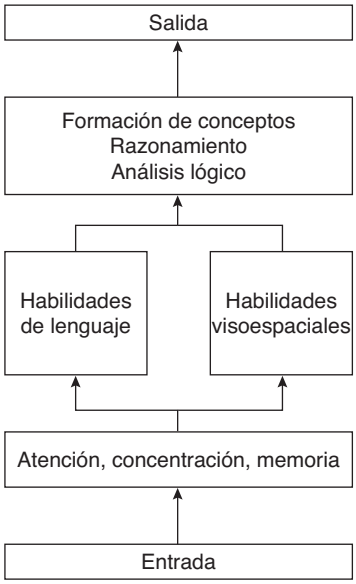


FIGURA 9.2 Modelo Reitan-Wolfson del funcionamiento neuropsicológico
Vea la explicación en el texto.
(Reproducido con autorización de R. M. Reitan.)

en el procesamiento central, la *fase de registro*, la cual consiste en la alerta, atención, observación continua y detección de la información que llega contra el telón de fondo de la experiencia previa. El proceso de detección involucra a las memorias inmediata, intermedia y de largo plazo. El registro de la información sensorial que llega es seguido por el procesamiento de la información verbal en el hemisferio izquierdo y de la información visual-espacial en el hemisferio derecho. El siguiente nivel superior en el procesamiento central consiste en la formación de conceptos, razonamiento y análisis lógico, funciones que generalmente tienen lugar por toda la corteza cerebral. La etapa final del modelo de Reitan-Wolfson es la salida —acciones motrices verbales y no verbales que resultan del procesamiento cognoscitivo de la entrada sensorial.

Etiología y sintomatología

Los trastornos neuropsicológicos pueden ser causados por anomalías genéticas, de desarrollo, envejecimiento o por trauma, tumores, abuso crónico del alcohol, dieta, drogas, microorganismos u otras condiciones físicas o químicas que afectan el funcionamiento del cerebro. Esos trastornos pueden afectar la atención, las habilidades motrices, habilidades visoespaciales, la memoria a corto y a largo plazos, el lenguaje y habilidades de pensamiento abstracto. También causan que el individuo se vuelva hiperactivo, impulsivo, fácil de distraer y emocionalmente inestable.

Cuando las áreas cerebrales del lenguaje están afectadas, pueden presentarse dificultades para entender el lenguaje hablado o escrito (*afasia*), así como deterioros en la habilidad para la lectura (*alexia*) y la escritura (*agrafia*). La *agnosia*, dificultad para reconocer objetos, puede ocurrir cuando se afectan las áreas sensoriales del cerebro. Y cuando se lesionan áreas motoras puede presentarse *apraxia*, la incapacidad para realizar movimientos propositivos, falta de coordinación e incluso parálisis.

Trastornos en los niños. Aunque existen múltiples causas de daño cerebral en todas las edades, los problemas que se derivan de la exposición prenatal al alcohol, las drogas y otros teratógenos, complicaciones durante el embarazo y el parto, y otros problemas del desarrollo temprano son causas comunes de daño cerebral en los niños pequeños. Rara vez resulta sencillo determinar las causas precisas de trastornos neuropsicológicos particulares en los niños debido a que ellos pasan por muchos otros cambios en esta época de la vida, y a que intentan adaptarse a muchas experiencias y acontecimientos nuevos. Además, los niños suelen ser menos cooperativos que los adultos durante los exámenes, y sus síntomas a menudo son más variables que los de los adultos.

Trastornos en los adultos mayores. Dos de los trastornos neuropsicológicos más relevantes en los adultos mayores son la demencia vascular y la enfermedad de Alzheimer. Los síntomas de esos trastornos incluyen confusión mental, pérdida de memoria, habla incoherente, mala orientación en el ambiente y, en algunos casos, falta de coordinación motriz, agitación, depresión y delirio. Los síntomas se vuelven más evidentes después de los 65 años, su frecuencia alcanza un punto máximo alrededor de los 70 años y después declina un poco. Los síntomas están asociados con degeneración neuronal, lo cual conduce a la atrofia (encogimiento) y a cambios degenerativos relacionados en el cerebro. El encogimiento ocurre sobre todo en la corteza frontal, la corteza temporal y la materia blanca asociada y puede reducir el cerebro de 15 a 30% de su peso previo.

En los años recientes se ha incrementado el uso de pruebas neuropsicológicas con el propósito de hacer diagnósticos diferenciales de pérdidas de memoria causadas por demencia, delirio y depresión. Los psicólogos que se especializan en el diagnóstico y tratamiento de los adultos mayores y en la investigación sobre este grupo de edad emplean muchas pruebas de este tipo. Por ejemplo, las pruebas de memoria, capacidades perceptuales y razonamiento abstracto se utilizan para diferenciar entre la demencia y la pseudodemencia de la depresión.

Pruebas neuropsicológicas

En años recientes, los avances tecnológicos en la imagenología cerebral (exámenes CT, MRI y PET) y otras técnicas de diagnóstico cerebal han sido impresionantes, pero el lugar, la extensión y los efectos del daño cerebral rara vez se identifican por completo sólo con procedimientos no psicológicos. Específicamente con propósitos de detección neuropsicológica, diagnóstico clínico detallado y planeación de intervención profesional, es que se han diseñado pruebas neuropsicológicas de sensación, velocidad y fuerza motriz, percepción e integración perceptual-motriz, lenguaje, atención, capacidad de abstracción, orientación y memoria. En la tabla 9.1 se presentan ejemplos de pruebas específicas que se aplican para evaluar funciones cognoscitivas y conductuales específicas que pueden ser afectadas por los trastornos neurológicos. Además de proporcionar una base para el tratamiento o la intervención profesional, los resultados obtenidos al aplicar pruebas neuropsicológicas contribuyen a la determinación de discapacidad en reclamaciones por accidentes ocupacionales, adjudicación de pensiones y otras circunstancias que implican compensación financiera. Las pruebas neuropsicológicas también se aplican en evaluaciones del

TABLA 9.1 Ejemplos de pruebas para evaluar déficit en ciertas funciones neuropsicológicas

Atención
WAIS-III Subprueba de retención de dígitos
WMS-III Retención espacial

Funciones ejecutivas: *habilidad de abstracción*
Prueba de categorías
WAIS-III Subprueba de semejanzas
Prueba Wisconsin de Clasificación de Tarjetas

Lenguaje
Evaluación Boston de Afasia Severa
Examen Boston de Diagnóstico de Afasia
WAIS-III Subprueba de vocabulario

Funciones de aprendizaje y memoria
Test Benton de Retención Visual
Escala de Memoria para Niños
Prueba Rey de Aprendizaje Verbal Auditivo
Escala de Memoria de Wechsler III
Evaluación de Rango Amplio de la Memoria y el Aprendizaje

Habilidades Visoespaciales
WAIS III Subprueba de diseño con cubos
Prueba de Figura Compleja y Reconocimiento

Funcionamiento intelectual global
Escala de Inteligencia para Adultos de Wechsler III
Escala de Inteligencia para Niños de Wechsler III
Woodcock-Johnson III

Instrumentos de detección
Prueba Rápida de Detección Neurológica II
Prueba de Detección para la Batería Neuropsicológica de Luria-Nebraska
Prueba Stroop de Detección Neuropsicológica

Funciones emocional-conductuales
Inventario Beck de Depresión
Lista de Verificación de la Conducta Infantil
Escala Hamilton de Depresión
Inventario Multifásico de Personalidad de Minnesota II
Inventario de Personalidad para Niños

Aprovechamiento académico
Prueba de Aprovechamiento Individual de Wechsler, segunda edición
Prueba de Aprovechamiento de Rango Amplio 3

Adaptado en parte de la Tabla 1 (p. 425) de Delis y Jacobson, 2000.

estado mental que contribuyen a tomar decisiones relacionadas con asuntos como la determinación de competencia, responsabilidad, demencia y otros asuntos legales.

La adquisición de competencia en la aplicación de las pruebas apropiadas y en el diagnóstico y tratamiento de los déficit en las capacidades neuropsicológicas requiere un largo programa de entrenamiento y experiencia intensiva. Incluso entonces, el diagnóstico y la intervención en materia neuropsicológica tienen tanto de arte como de ciencia y son procesos sujetos a numerosos escollos. Para un diagnóstico comprensivo, la información obtenida de la aplicación de pruebas neuropsicológicas debe ser complementada con una historia de caso detallada, observaciones cuidadosas, calificaciones de la conducta del paciente y varias pruebas médicas.

Para evaluar los efectos del trauma o de otras causas de lesión al cerebro es importante obtener un estimado del funcionamiento cognoscitivo premórbido del sujeto. Esto puede lograrse de varias maneras, quizá con mayor precisión de las calificaciones obtenidas en pruebas estandarizadas de inteligencia o aprovechamiento aplicadas antes de que ocurriera la lesión. Otros indicadores del funcionamiento premórbido, aunque menos precisos, son el nivel educativo y la posición socioeconómica. Además, debe tenerse en mente que las funciones en diferentes áreas del cerebro varían no sólo con su localización, sino también con la edad cronológica, el género y otros factores demográficos.

La WCST y otras pruebas de detección. Dado que un examen neuropsicológico completo es un proceso que consume tiempo, se ha elaborado una serie de pruebas de detección cortas como preliminares a la aplicación de una batería más extensa. Algunos ejemplos son: Examen Cog-

noscitivo Neuropsicológico Breve, Prueba Rápida de Detección Neurológica, Detector Bayley del Neurodesarrollo Infantil, Prueba de Detección para la Batería Neuropsicológica Luria-Nebraska, Prueba Stroop de Detección Neuropsicológica y Prueba Wisconsin de Clasificación de Tarjetas. La última es quizá la que se aplica con mayor frecuencia y la más investigada de todos los instrumentos de detección neuropsicológica.

La Prueba Wisconsin de Clasificación de Tarjetas (WCST) (de PAR) evalúa la perseveración y el pensamiento abstracto. Es sensible en particular a la disfunción del lóbulo frontal y útil para diferenciar entre lesiones frontales y no frontales. No se cronometra (20 a 30 minutos) y es apropiada para un rango amplio de edad (de 6.5 a 80 años). La WCST consta de cuatro tarjetas de estímulo y un paquete de 64 tarjetas de respuesta. Cada tarjeta de respuesta contiene uno de cuatro símbolos (triángulo, estrella, cruz o círculo) en uno de cuatro colores (rojo, verde, amarillo o azul). Se indica al examinado que clasifique las tarjetas de respuesta por debajo de las cuatro tarjetas de estímulo de acuerdo con cierto principio (color, forma o número). No se informa al examinado del principio de clasificación, sino sólo si sus respuestas son correctas o equivocadas. Después de que se han dado diez respuestas correctas consecutivas, el examinador cambia el principio de clasificación sin advertencia (digamos de "color" a "forma"). La calificación suele hacerse en términos del número de ensayos necesarios para dar un cierto número de respuestas correctas consecutivas usando cada principio de clasificación. Tanto las 64 tarjetas como versiones para computadora de la WCST se encuentran disponibles en Psychological Assessment Resources. El manual revisado proporciona información normativa, de confiabilidad y de validez de la prueba, que está basada en muestras de niños y adolescentes. Sin embargo, Egeland (1985) recomendó cautela al usar esta prueba con propósitos clínicos, y Mountain y Snow (1993) cuestionaron su sensibilidad diferencial al daño del lóbulo frontal.

WAIS-R y WAIS-III como pruebas neuropsicológicas. Los cambios en la habilidad mental general que resultan de trastornos neuropsicológicos pueden ser detectados mediante la aplicación de pruebas de inteligencia como la WAIS-R, la WAIS-III y la WISC-III. Diferencias significativas (de 10 puntos o más) observadas en esas pruebas entre las calificaciones en los CI verbal y de desempeño, además de una dispersión pronunciada de la calificación escalada de subprueba, pueden ser indicadores de trastorno cerebral traumático e incluso proporcionar indicios sobre la localización del daño nervioso. Un CI verbal significativamente menor al CI de desempeño sugiere un daño bien definido en el hemisferio izquierdo, mientras que un CI de desempeño significativamente menor que el CI verbal sugiere un daño bien definido en el hemisferio derecho. Sin embargo, un desempeño significativamente inferior al CI verbal también se asocia con daño difuso del cerebro.

La necesidad de obtener una definición más clara de los efectos del daño cerebral orgánico en el funcionamiento cognoscitivo y conductual llevó al desarrollo de una modificación de la WAIS-R denominada WAIS-R como Instrumento Neuropsicológico (WAIS-R NI). A excepción de algunas modificaciones, como en los rompecabezas del Ensamble de Objetos, las subpruebas de la WAIS-R fueron conservadas en la WAIS-R NI. Además, se proporcionaron las siguientes subpruebas: Información Opción Múltiple, Vocabulario Opción Múltiple, Aritmética Lápiz y Papel, Semejanzas Opción Múltiple, Ordenamiento de Frases, Retención Espacial y Copia de Símbolos. Es posible obtener una mejor evaluación de las funciones cognoscitivas deterioradas y no deterioradas comparando los resultados obtenidos del foco en la memoria de recuerdo de las subpruebas convencionales con el foco en la memoria de reconocimiento de las nuevas subpruebas y los procedimientos convencionales de aplicación de las viejas subpruebas con los procedimientos alternativos de aplicación de las nuevas subpruebas. Además de las comparaciones

de calificaciones, un análisis de los errores y las estrategias empleadas por los examinados arroja información útil para el diagnóstico y la rehabilitación.

Pruebas perceptivas-memoria. La observación de que en el caso de daño cerebral ocurren distorsiones en la percepción y la memoria llevó al desarrollo de pruebas especiales de diagnóstico como el Test Gestáltico Visomotor de Bender (WPA) y el Test de Benton de Retención Visual (The Psychological Corporation). Esas dos pruebas se administran con frecuencia como complemento a pruebas individuales de inteligencia y a otros exámenes psicológicos.

El Test Gestáltico Visomotor de Bender consta de 9 diseños geométricos en tarjetas blancas, de 4×6 pulgadas, las cuales se muestran una a la vez al examinado y se le pide que las copie. Las distorsiones significativas en el copiado de los diseños se interpretan como déficit en la percepción. Los niños de ocho años y mayores de inteligencia promedio o superior al promedio, por lo general, no cometen más de dos errores en la prueba Bender. Los errores que se consideran indicadores de daño cerebral orgánico incluyen distorsiones de forma; rotación del diseño; problemas para integrar el diseño; dibujos desproporcionados, traslapados o fragmentados; y perseveraciones (Lacks, 1984).

El Test de Benton de Retención Visual consiste en diez diseños presentados de manera individual al examinado. A diferencia del Bender, en el cual el examinado hace un dibujo mientras mira la tarjeta correspondiente, en el Benton se muestra al examinado cada diseño y luego él trata de copiarlo de memoria. Las formas pequeñas incluidas en la periferia de la mayoría de los dibujos se consideran importantes para determinar la habilidad del examinado para mantener la integridad del campo visual. El Benton se califica, al igual que el Bender, de acuerdo con el número y tipo de errores. La investigación con el test de Benton ha proporcionado apoyo a su sensibilidad al daño cerebral traumático, al trastorno por déficit de atención y a varios tipos de demencia.

Déficit de memoria y pruebas. Los problemas con la memoria de corto y largo plazos no sólo son indicadores de retraso mental, sino de discapacidades específicas de aprendizaje, trauma cerebral, trastornos neurológicos, trastorno por déficit de atención con hiperactividad (TDAH), envejecimiento e incluso trastornos emocionales. Las deficiencias en la memoria de recuerdo, y en particular el recuerdo libre, son más pronunciadas que los deterioros en la memoria de reconocimiento en las personas con daño cerebral. Por lo regular, los pacientes muestran menos déficit en las pruebas de reconocimiento o memoria de identificación que en las de memoria de recuerdo, menos déficit en la memoria implícita que en la explícita, y menos déficit en la memoria de habilidades que en la de acontecimientos. Debido a que las pruebas individuales de inteligencia como las de la serie Wechsler generalmente enfatizan el recuerdo libre, los pacientes con lesiones cerebrales pueden estar en mayor desventaja y aparecer más dañados en esas pruebas.

Dado que la memoria y el aprendizaje no son habilidades unitarias, a menudo se necesita una batería de pruebas para identificar la presencia de déficit específicos. Dichas baterías no pueden tomar el lugar de las pruebas de inteligencia, las cuales evalúan un rango más amplio de funciones cognoscitivas, pero pueden proporcionar datos complementarios e indicios para el diagnóstico. Cuatro baterías populares para la evaluación de la memoria son la Escala de Memoria de Wechsler, tercera edición (WMS-III), la Prueba de Memoria y Aprendizaje (TOMAL), la Evaluación de Rango Amplio de la Memoria y el Aprendizaje (WRAML) y las Escalas de Evaluación de la Memoria (MAS). La WMS-III mide la memoria para estímulos auditivo-verbales y visuales-no verbales, material significativo y abstracto, para modos de recuerdo inmediato y

demorado, en individuos de 16 a 89 años. La TOMAL (pro.ed), la WRAML (Wide Range) y la MAS (The Psychological Corporation) miden funciones de memoria verbal y no verbal (visual). Las dos primeras pruebas están diseñadas para niños y adolescentes y la última para adultos. Las tres baterías son medidas de alta confiabilidad para las funciones de memoria y aprendizaje. El manual de la MAS proporciona perfiles de calificaciones para pacientes con trastornos neurológicos como la demencia, daño interno de la cabeza, y lesiones de los hemisferios izquierdo y derecho.

Baterías de pruebas neuropsicológicas. Aunque las pruebas convencionales de inteligencia, como las de la serie Wechsler, son útiles para identificar déficit neuropsicológicos, por tradición se ha aplicado una batería de pruebas, como las que componen la Batería Halstead-Reitan de Pruebas Neuropsicológicas y la Batería Neuropsicológica de Luria-Nebraska, para medir las habilidades adaptativas de base neuropsicológica que no son evaluadas por las pruebas de inteligencia. Las respuestas a los materiales de esas baterías proporcionan información útil a los psicólogos a quienes se pide evaluar relaciones cerebro-conducta, proporcionar opiniones sobre la presencia de enfermedad o daño cerebral, planear programas de rehabilitación y dar testimonio legal concerniente a sus evaluaciones neuropsicológicas.

Batería Halstead-Reitan de Pruebas Neuropsicológicas. En la tabla 9.2 se describen los materiales que constituyen la Batería Compuesta de la Batería Halstead-Reitan de Pruebas Neuropsicológicas (Reitan Neuropsychology Laboratory). Diferentes formas de esas pruebas se incluyen en la Batería para Adultos (para edades de 15 años en adelante), la Batería para Niños Mayores (edades de 9 a 14 años) y la Batería para Niños Pequeños (edades de 5 a 8 años). Las pruebas y los procedimientos tocan una serie de habilidades sensoriales, velocidad y destreza perceptual-motriz, funciones de lenguaje expresivo y receptivo, memoria, formación de conceptos y razonamiento abstracto. Cualquiera de esas habilidades puede ser afectada por daño o disfunción del sistema nervioso central o de los receptores sensoriales y los músculos. Entre las pruebas más complejas de la Halstead-Reitan se encuentran la Prueba de Categorías y la Prueba de Trazo de Pistas. En la Prueba de Categorías el examinado deduce principios generales a partir de la información presentada en diapositivas. En la Prueba de Trazo de Pistas dibuja líneas que conectan círculos con números y letras (de 1 a A, de 2 a B, etc., alternando números y letras).

Batería Neuropsicológica de Luria-Nebraska. Esta batería de pruebas (de WPS) fue diseñada para evaluar: dominancia cerebral; funciones táctiles, visuales y motrices; percepción y reproducción de tonos y ritmo; habla receptiva y expresiva; lectura, escritura y aritmética; memoria; formación de conceptos y otros procesos intelectuales. Ambas formas (I y II) de la batería pueden calificarse por computadora, pero la Forma I también puede calificarse a mano. Al igual que la Halstead-Reitan, la Luria-Nebraska se administra para efectuar una detección neuropsicológica más pormenorizada del daño cerebral. La administración de la Luria-Nebraska sólo se lleva una tercera parte del tiempo requerido por la Halstead-Reitan, pero se le ha criticado por confiar demasiado en las habilidades de lenguaje y por no hacer una detección adecuada de la afasia y otros trastornos neuropsicológicos.

Evaluación neuropsicológica basada en la computadora. Los avances en la neurofisiología y la psicología cognoscitiva, junto con progresos en la tecnología de las computadoras y la metodología psicométrica durante las tres décadas pasadas, han conducido a un mayor uso de las

TABLA 9.2 Pruebas y procedimientos para la Batería Halstead-Reitan de Pruebas Neuropsicológicas

Prueba de categorías. Mide el razonamiento abstracto y la formación de conceptos; requiere que el examinado encuentre una regla para categorizar las ilustraciones de formas geométricas.

Prueba de desempeño táctil. Mide la habilidad cinestésica y sensoriomotriz; requiere que el examinado, con los ojos vendados, coloque cubos en lugares apropiados sobre un tablero vertical con la mano dominante, luego con la mano no dominante, después con ambas manos; también mide la memoria incidental de cubos.

Prueba de percepción de los sonidos del lenguaje. Mide la atención y la síntesis auditiva-visual; requiere que el examinado elija de entre cuatro opciones la versión escrita de palabras grabadas sin sentido.

Prueba del ritmo de Seashore. Mide la atención y la percepción auditiva; requiere que el examinado indique si ritmos musicales pareados son iguales o diferentes.

Prueba de golpes dactilares. Mide la velocidad motriz; requiere que el examinado golpee una palanca similar a la del telégrafo con tanta rapidez como sea posible por 10 segundos.

Fuerza de agarre. Mide la fuerza del agarre con un dinamómetro; requiere que el examinado apriete tan fuerte como sea posible; se hacen intentos separados con cada mano.

Trazo de pistas, *partes A y B*. Mide la habilidad para rastrear, la flexibilidad mental y la rapidez; requiere que el examinado, bajo presión de tiempo, conecte números (parte A) o números y letras en orden alternado (parte B) mediante una línea a lápiz.

Reconocimiento táctil de formas. Mide la habilidad sensorial-perceptual; requiere que el examinado reconozca formas simples (por ejemplo, triángulos) colocadas en la palma de la mano.

Examen sensorial-perceptual. Mide la habilidad sensorial-perceptual; requiere que el examinado responda a tareas sensoriales bilaterales simples, por ejemplo, detectar qué dedo ha sido tocado, qué oído ha recibido un sonido breve; evalúa los campos visuales.

Prueba de detección de afasia. Mide las habilidades de lenguaje expresivo y receptivo; las tareas incluyen nombrar un reactivo ilustrado (por ejemplo, un tenedor); repetir frases cortas. La tarea de copiado (que no es una medida de la afasia) se incluye aquí por razones históricas.

Complementarias. WAIS-III, WRAT-3, MMPI-2, pruebas de memoria como la Escala de Memoria Wechsler-III, o la Prueba Rey de Aprendizaje Verbal Auditivo.

Adaptado de Robert J. Gregory, *Psychological testing: History, principles, and applications* (tercera edición). Copyright © 2000 por Allyn & Bacon.

computadoras para administrar, calificar e interpretar las pruebas neuropsicológicas. Como resultado, la examinación neuropsicológica se ha vuelto más rápida, más flexible y más centrada; por medio de la evaluación basada en computadoras es posible determinar no sólo la exactitud de las respuestas, sino también su rapidez e incluso su intensidad.

Entre las muchas pruebas neuropsicológicas con versiones basadas en la computadora se encuentran la Prueba de Categorías y la Prueba Wisconsin de Clasificación de Tarjetas. También se dispone de software de computadora para los componentes de la Batería Neuropsicológica Halstead-Reitan y la Batería Neuropsicológica de Luria-Nebraska. Además de las pruebas sen-

cillas y de las baterías de pruebas que pueden aplicarse por un examinador en persona o por una computadora existen instrumentos que sólo son aplicados por medio de la computadora. Un ejemplo es MicroCog: Evaluación del Funcionamiento Cognoscitivo (The Psychological Corporation). Diseñado para evaluar el funcionamiento cognoscitivo en adultos de 18 a 80 años, MicroCog viene en una forma estándar que requiere de 50 a 60 minutos y en una forma breve para administrarse en 30 minutos. Las 18 pruebas de la forma normal fueron estandarizadas en 810 adultos de quienes se dijo eran representativos de la población nacional estadounidense, con normas separadas para nueve grupos de edad así como con normas ajustadas para el nivel educativo. Se proporcionan calificaciones resumidas para nueve áreas de funcionamiento: Atención/Control Mental, Memoria, Razonamiento/Cálculo, Procesamiento Espacial, Tiempo de Reacción, Precisión del Procesamiento de Información, Velocidad del Procesamiento de Información, Funcionamiento Cognoscitivo y Competencia Cognoscitiva. En el manual se proporcionan datos de validez para varios grupos clínicos (depresión mayor, demencia, esquizofrenia, alcoholismo, epilepsia, psiquiátrico mixto, lupus y otros) y correlaciones con otras pruebas neuropsicológicas.

RESUMEN

Los estudios de desarrollo humano en la infancia y la niñez temprana, entre los cuales destacan los conducidos por Arnold Gesell y sus colegas en la Universidad de Yale durante las décadas de 1920 y 1930, proporcionaron normas del desarrollo y pruebas que han servido como directrices y métodos para la práctica y la investigación con niños.

Los Programas de Desarrollo de Gesell, las Escalas de Bayley de Desarrollo Infantil (BSID-II), la Prueba Denver de Detección del Desarrollo (Denver II) y otras medidas de habilidades en infantes y niños pequeños han contribuido al conocimiento científico del desarrollo y los trastornos de la niñez. Por desgracia, las pruebas de inteligencia infantil no tienen alta confiabilidad ni proporcionan una buena predicción del desarrollo y el desempeño cognoscitivo posterior. Las tareas sensoriomotrices en las pruebas infantiles, combinadas con la falta de atención y la baja motivación de los jóvenes examinados, contribuyen a las bajas correlaciones entre las calificaciones en las pruebas presentadas durante los dos o tres primeros años de vida y las calificaciones obtenidas por los mismos niños en la edad escolar.

La mayoría de las pruebas diseñadas para evaluar y seguir las demoras en el desarrollo de los niños pequeños siguen en la actualidad los criterios especificados en el Acta para la Educación de Individuos con Discapacidades (IDEA) estadounidense. Dos ejemplos de pruebas diseñadas de manera específica de acuerdo con los cinco dominios de la IDEA son la Prueba de Detección FirstSTEP para la Evaluación de Preescolares y los Perfiles de Detección Temprana AGS. Una prueba antigua, pero todavía de gran uso que se adhiere de manera cercana a los criterios de la IDEA son las Escalas McCarthy de las Habilidades de los Niños.

Los trastornos específicos de aprendizaje son las discapacidades en la lectura, escritura, ortografía, aritmética u otras habilidades académicas que no pueden ser explicadas por el retraso mental, impedimentos sensorial-motrices específicos, trastornos emocionales o desventajas ambientales. La ley federal estadounidense ordena que los niños con discapacidades deben ser diagnosticados de manera profesional y que debe prepararse un plan individualizado de educación para cada niño. En la mayoría de los estados un indicador psicodiagnóstico importante de una discapacidad de aprendizaje es cuando la calificación de un niño en una prueba estandariza-

da de aprovechamiento es significativamente inferior a su calificación en una prueba co-normada de inteligencia. Entre las diversas pruebas de inteligencia y aprovechamiento que han sido empleadas en la determinación de discapacidades específicas de aprendizaje están la WISC-III y las Pruebas Wechsler de Aprovechamiento Individual, la K-ABC y la Prueba Kaufman de Aprovechamiento Educativo, así como las pruebas de habilidades cognoscitivas y de aprovechamiento en la Woodcock-Johnson III.

Los médicos y los psicólogos emplean varias técnicas y procedimientos en un intento por entender las causas y consecuencias de los trastornos cerebrales, y para hacer recomendaciones sobre el tratamiento apropiado. Las observaciones conductuales, entrevistas con el paciente y con otras personas, pruebas neurológicas, procedimientos de imagenología cerebral y pruebas psicológicas pueden contribuir a lograr el diagnóstico, la planeación del tratamiento y el pronóstico de los trastornos neuropsicológicos.

Las diferencias entre las calificaciones escaladas en las diversas subpruebas de la Escala de Inteligencia para Adultos de Wechsler-III (WAIS-III) y las Escalas de Inteligencia para Niños de Wechsler-III (WISC-III), así como las diferencias entre los CI verbal y de EJECUCIÓN de esos instrumentos, pueden proporcionar información sobre la localización y gravedad del trastorno neuropsicológico. También están disponibles muchas pruebas especializadas de funcionamiento neuropsicológico. Entre esas pruebas se incluyen medidas perceptuales-memoria como el Test Gestáltico Visomotor de Bender y el Test de Benton de Retención Visual, pruebas breves de detección como la Prueba Wisconsin de Clasificación de Tarjetas, pruebas de memoria a corto y largo plazos como la Escala de Memoria Wechsler, la Prueba de Memoria y Aprendizaje, la Evaluación de Rango Amplio de la Memoria y el Aprendizaje y las Escalas de Evaluación de la Memoria. Para un análisis y un diagnóstico más comprensivos de un trastorno neuropsicológico se recomienda la aplicación de una batería completa de pruebas (por ejemplo, la Batería de Pruebas Neuropsicológicas Halstead-Reitan, la Batería Neuropsicológica de Luria-Nebraska). Muchas pruebas neuropsicológicas contemporáneas pueden ser aplicadas por un examinador en persona o por una computadora. MicroCog, un instrumento de detección para adultos con deterioro cognoscitivo de leve a moderado, se aplica exclusivamente por computadora.

PREGUNTAS Y ACTIVIDADES

1. Defina cada uno de los siguientes términos:

TDAH	dislexia
agnosia	IDEA
trastorno de Alzheimer	PEI
afasia	prueba de inteligencia infantil
apraxia	discapacidad de aprendizaje
niños en riesgo	trastorno neuropsicológico
ataxia	prueba neuropsicológica
área de Broca	plano temporal
edad de desarrollo (ED)	modelo Reitan-Wolfson
cociente de desarrollo (CD)	prueba de detección
discalculia	área de Wernicke

2. ¿Qué tan grandes son las correlaciones entre las calificaciones en pruebas aplicadas a infantes y niños muy pequeños y las calificaciones obtenidas por los mismos niños en pruebas de inteligencia

presentadas a una edad posterior? ¿Cómo son influidas las magnitudes de esas correlaciones por las confiabilidades de las pruebas de inteligencia para infantes y niños pequeños, por el hecho de que las pruebas miden factores diferentes que las pruebas de inteligencia aplicadas más tarde en la niñez, y las diferencias en los procedimientos para aplicar pruebas a los infantes y niños pequeños y las pruebas aplicadas a los niños mayores?

3. Obtenga tanta información como pueda sobre las previsiones de varias leyes concernientes a la identificación, el diagnóstico y los programas de intervención con niños que están médica y/o ambientalmente en riesgo. Concéntrese en las leyes públicas estadounidenses PL 94-142, PL 99-457, PL 101-476 (IDEA), PL 101-336 (ADA) y PL 102-119.

4. ¿Qué características y habilidades piensa que contribuyen a ser un examinador psicológico efectivo de infantes y niños pequeños? ¿En qué se asemejan y en qué se distinguen esas características y habilidades con las requeridas para probar a niños mayores y adultos?

5. Compare FirstSTEP con los Perfiles de Detección Temprana AGS, y compare DAYC con la Evaluación del Desarrollo de Infantes y Niños Pequeños en términos de sus propósitos, composición y calificación.

6. Mencione y describa varias causas y tipos de discapacidades de aprendizaje y los tipos de procedimientos de intervención que puede esperarse mejoren esas condiciones.

7. Debido a que en Estados Unidos los criterios con los que se diagnostican las discapacidades de aprendizaje, incluyendo el procedimiento estadístico para determinar la discrepancia entre habilidad y aprovechamiento, varían de un estado a otro, ¿sería posible que un niño tuviera una discapacidad de aprendizaje en un estado y no en otro? ¿Qué consecuencias puede tener esto para el niño, para los gobiernos y para la población de los estados?

8. Mencione una o dos pruebas de memoria y una o dos pruebas de habilidades perceptivo-motrices y los propósitos para los cuales podrían utilizarse.

9. Mencione las conductas y los síntomas cognoscitivos de varios trastornos neuropsicológicos descritos en libros sobre neuropsicología, psicología fisiológica y psicología anormal. ¿Qué contribuciones pueden hacer los psicólogos para el diagnóstico y tratamiento de dichos trastornos?

CAPÍTULO DIEZ

EVALUACIÓN DE HABILIDADES ESPECIALES

El término *aptitud* se ha definido tradicionalmente como la habilidad de aprovechar la educación o capacitación obtenida en un campo determinado, mientras que *aprovechamiento* se refiere al grado de habilidad ya obtenida. La medida de la aptitud se centra en el futuro, la del aprovechamiento en el pasado. Así, las pruebas de aptitud se han diseñado sobre todo para evaluar el aprovechamiento potencial o para predecir el desempeño futuro en algún campo o intento.

Las habilidades de una persona se evalúan con fines de asesoría y colocación académica y laboral. Con información sobre una prueba de aptitud en mano, los asesores o jefes de personal pueden mejorar su trabajo al aconsejar a las personas o ubicarlas en los programas apropiados de educación y capacitación o en puestos de trabajo adecuados.

CONCEPTOS Y CARACTERÍSTICAS DE LAS HABILIDADES ESPECIALES

En cierto sentido, el término *aptitud* no es acertado si se pretende interpretar como una característica innata, inmutable por medirse. Los primeros evaluadores de la mente aspiraban a medir características hereditarias, pues suponían que todas las personas que examinaban tenían las mismas oportunidades para aprovechar las experiencias en que se basaban los materiales de prueba. No obstante, esta suposición era incorrecta: las experiencias, y por ende las oportunidades para aprender, nunca son exactamente iguales para personas distintas, sobre todo si las personas provienen de clases sociales o culturas diferentes.

En la actualidad, generalmente se reconoce que las pruebas de aptitud son medidas de aprovechamiento, un producto complejo de la interacción entre influencias hereditarias y ambientales. A la inversa, si las *pruebas de aptitud* son instrumentos psicométricos que pueden predecir el logro futuro, entonces las pruebas de aprovechamiento que prefiguran las notas escolares y otros criterios también califican como medidas de habilidad.

Debido a la confusión sobre la diferencia entre aptitud y logro, se ha recomendado que ambos términos se reemplacen con el término único de *habilidad*. Entonces, dependiendo del propósito para el que se utilice —evaluar el conocimiento y la comprensión presentes o pronosticar el desempeño futuro—, una prueba de habilidad puede ser tanto una medida de aprovechamiento como de aptitud. Pero podría ser un error suponer que la distinción entre aptitud y aprovechamiento carece de consecuencias. Como ejemplo de la diferencia funcional entre medidas de estas dos variables, considérense los resultados de un estudio realizado por Carroll (1973). Se descubrió que el desempeño en un curso de lengua extranjera para estudiantes cuyas calificaciones anteriores al curso en una prueba de aprovechamiento de lengua fueron de cero, podría predecirse a partir de las calificaciones en una prueba de aptitud para aprender lenguas

extranjeras. Al final del curso ambas pruebas se aplicaron de nuevo. Como podría esperarse, si la capacitación había mejorado el aprovechamiento sin alterar la aptitud, las calificaciones de la prueba de aprovechamiento aumentaron considerablemente, pero las calificaciones en las pruebas de aptitud permanecieron sin modificaciones en lo esencial.

Habilidades generales y específicas

Las pruebas de inteligencia analizadas en los capítulos 7, 8 y 9 son medidas de aptitud *general*, en cuanto a que las calificaciones en esas pruebas representan un compuesto de habilidades cognoscitivas que puede usarse para predecir el aprovechamiento y otros comportamientos en un amplio espectro de situaciones. De hecho, las calificaciones en las pruebas de inteligencia general a menudo son mejores para predecir el éxito en situaciones educativas y laborales que las calificaciones combinadas en medidas de habilidades especiales. Pero el que las pruebas de inteligencia general midan una mezcolanza de aptitudes o habilidades específicas es una espada de dos filos. En este hecho radican tanto las ventajas como las deficiencias de estos tipos de pruebas.

Debido a que las pruebas de inteligencia miden una combinación de habilidades, tienen lo que Cronbach (1970) llamaba una extensa *amplitud de banda*. Una ventaja de su amplio contenido es que las pruebas de inteligencia son moderadamente eficaces para pronosticar un amplio espectro de criterios de desempeño. Una prueba más extensa de sólo una de las habilidades especiales medidas por una prueba de inteligencia, es decir, un instrumento con una amplitud de banda más angosta, al parecer tiene una mayor *fidelidad*. En otras palabras, se esperaría que midiera una variable específica con mayor precisión y predijera mejor un espectro menor de criterios.

Al observar las correlaciones positivas significativas entre las medidas de habilidades, Vernon (1960) concluyó que la inteligencia general es más importante que las habilidades especiales para determinar el éxito laboral (vea también Hunter y Schmidt, 1996). En la medida en que esto sea cierto, probablemente se deba a que los criterios del éxito laboral, como las calificaciones de las pruebas de inteligencia, son productos complejos de múltiples variables. En otras palabras, los criterios laborales tienen una amplitud de banda extensa y, por lo tanto, mayor probabilidad de predecirse con más precisión a partir de una combinación de medidas, más que por calificaciones de una única prueba de habilidades especiales.

Orígenes de la evaluación vocacional

Un acontecimiento que impulsó el desarrollo de pruebas de habilidades especiales durante las décadas de 1920 y 1930 fue el crecimiento de la administración científica. Quienes promovían la administración científica en los negocios y la industria consideraban que tanto los empleados como los jefes se beneficiarían con el diseño de pruebas psicológicas que pudieran contribuir a conjuntar personas y puestos de trabajo. Sostenían que el uso de pruebas daría como resultado que se eligieran empleados para, y se ubicaran en, los empleos que pudieran desempeñar con mayor eficiencia. Seleccionar empleados más competentes y asignarles puestos para los que fuesen más aptos, incrementando la productividad, beneficiaría tanto a empleados como a empleadores, y a la organización en su totalidad.

Durante los años de la Gran Depresión en la década de 1930, cuando los asuntos relacionados con el empleo eran de particular interés para el gobierno, los programas de investigación y desarrollo en la Universidad de Minnesota, y en otros sitios, dieron origen a la construcción de una serie de pruebas de habilidades especiales para usarse en consejería vocacional y selección y colocación de empleados. A partir de estos programas y de subsecuentes esfuerzos se crearon no sólo numerosas medidas de habilidades individuales, sino también varias baterías de pruebas.

Validez de las pruebas de habilidades especiales

Debido a que las pruebas de aptitudes o de habilidades especiales se diseñan teniendo en mente una predicción diferencial, es razonable preguntarse qué tanto lograrán predecir quién tendrá éxito y quién fracasará en ocupaciones o programas de capacitación particulares. Es decir, ¿exactamente qué tan válidas son las pruebas de aptitud vocacional? La respuesta es que, en general, la validez de estas pruebas no es muy elevada. Como se muestra en la tabla 10.1, los coeficientes de validez promedio de diferentes tipos de pruebas de aptitud para predecir el desempeño en varias categorías de empleos suelen encontrarse en los .20 y casi nunca por encima de los .30 (Ghiselli, 1973). Estos coeficientes tan modestos dan puntuaciones más bajas a las limitaciones de estos tipos de pruebas para predecir el desempeño en el trabajo.

Además de las características de las pruebas mismas, la validez resulta alterada por problemas en los criterios para especificar y medir el éxito en el trabajo. Los acontecimientos incidentales o fortuitos que no se previeron, tales como los cambios económicos y sociales que influyen en la situación del empleo, pueden afectar la validez predictiva de las pruebas en contextos institucionales o industriales. A pesar de sus limitaciones, tales pruebas todavía pueden colaborar en la determinación de la ocupación o el programa de capacitación más adecuado para una persona dada. Ciertamente, las pruebas son limitadas cuando se usan solas, pero su valor aumenta cuando se combinan las calificaciones con otro tipo de información (intereses, motivación, actitudes y cuestiones similares) sobre las personas.

Ni siquiera los coeficientes de validez relativamente modestos de la mayoría de las pruebas de habilidades especiales están establecidos: varían con el carácter del criterio, la situación y las personas examinadas. Por ejemplo, es probable que un coeficiente de validez sea más alto cuando en un programa de capacitación se valida una prueba contra los grados alcanzados que cuando se valida contra las tasas del desempeño real en el trabajo (vea la tabla 10.1). Un coeficiente de validez también tiende a ser más alto cuando la prueba se administra y los datos del criterio se unifican en un lapso bastante breve que cuando hay una gran demora entre la aplicación de la prueba y la unificación de los datos de criterio.

TABLA 10.1 Correlaciones entre pruebas de habilidad y criterios ocupacionales

	HABILIDAD							
	INTELECTUAL		ESPACIAL Y MECÁNICA		PRECISIÓN PERCEPTUAL		MOTRIZ	
CATEGORÍA OCUPACIONAL	*Capacit.**	*Dest.*	*Capacit.*	*Dest.*	*Capacit.*	*Dest.*	*Capacit.*	*Dest.*
Empleados	.47	.28	.34	.17	.40	.29	.14	.16
Gerentes	.30	.27	.28	.22	.23	.25	.02	.14
Protección	.42	.22	.35	.18	.30	.21		.14
Ventas		.19		.18		.04		.12
Servicios	.42	.27	.31	.13	.25	.10	.21	.15
Comercio y artesanías	.41	.25	.41	.23	.35	.24	.20	.19
Op. de vehículos	.18	.16	.31	.20	.09	.17	.31	.25

*Capacit., criterios de capacitación; Dest., criterios de destreza.

Fuente: Con base en datos de Ghiselli, 1973.

Se advierte variabilidad situacional en el coeficiente de validez cuando la correlación entre calificaciones en una prueba de habilidad y las tasas del desempeño son menores en una organización que en otra. Sin embargo, la investigación ha demostrado que muchas pruebas de selección de empleos tienen una gran cantidad de *generalización de validez*; es decir, son válidas a través de una amplia gama de situaciones (vea, por ejemplo Hunter y Schmidt, 1990; Schmidt *et al.*, 1993; Schmidt, Ones y Hunter, 1992). Sin embargo, la validez de una prueba varía en cierto grado según la situación y las características de las personas que se examinan. Puede variar de acuerdo con el sexo, la etnia y la condición socioeconómica de los examinados, así como con sus intereses vocacionales, motivación y característica de personalidad. Tales diferencias individuales y por grupo, que influyen o moderan la correlación entre una prueba y una medida de criterio, se denominan *variables moderadoras*.

Las instituciones son como las personas en cuanto a que sienten la motivación no sólo de sobrevivir, sino de crecer; de hecho, en nuestra dinámica sociedad las instituciones deben expandirse o, a largo plazo, fracasarán. En consecuencia, desde la perspectiva de la institución, un factor importante al decidir usar una prueba específica para seleccionar, ubicar o promover al personal es determinar si la prueba contribuye al bienestar económico de la organización. El costo de aplicar la prueba debe sopesarse contra los beneficios que se obtendrán al usarla, y los estudios sobre la validez de la prueba pueden ayudar a medir estos beneficios. Una prueba no sólo debe ser un pronosticador eficiente y válido del desempeño en el trabajo, también deberá ser un pronosticador independientemente válido. ¿Por qué usar la prueba si hay disponibles métodos más baratos de identificar buenos trabajadores y de prever cómo se desempeñarán?

El beneficio económico para la institución no es, desde luego, la única razón que lleva a determinar la validez de una prueba para un fin específico en una situación determinada. Una razón legal importante que justifica llevar a cabo estudios de validez en las empresas y la industria se concentra en el problema del sesgo o la justicia. Por ejemplo, puede ser que la correlación entre prueba y criterio sea considerablemente más alta con un grupo étnico o de género que con otro. De ser así, es injusto usar la misma ecuación de predicción con ambos grupos. La *justicia*, o relativa carencia de sesgo, de la prueba tiene que demostrarse si se planea usarla con propósitos de selección o de clasificación.

El desempeño y las pruebas de lápiz y papel

Las primeras pruebas de habilidades especiales eran pruebas de desempeño que demandaban a los examinados construir algo o manipular objetos físicos de una forma determinada. Tales pruebas de *aparatos* con frecuencia son más interesantes que las pruebas de lápiz y papel, en especial para los examinados con problemas de lectura. Pero las confiabilidades de las pruebas de desempeño de velocidad suelen ser menores que las de las pruebas comparables de lápiz y papel, y las pruebas requieren de mucho tiempo y son costosas de aplicar. Asimismo, las correlaciones entre calificaciones en las pruebas de desempeño y las medidas de lápiz y papel para la misma habilidad están lejos de ser perfectas. A pesar de las desventajas de las pruebas de desempeño, las *pruebas de muestras de trabajo* (o *pruebas de réplica del empleo*), que demandan a los examinados realizar una muestra de tareas similares a las que comprende determinado trabajo, se encuentran entre las medidas de habilidad más útiles en contextos ocupacionales específicos.

Técnica de charola de pendientes y centros de evaluación

Un ejemplo interesante de una prueba de muestras de trabajo es la *técnica de charola de pendientes*. Este procedimiento se diseñó originalmente para evaluar al personal administrativo de

las escuelas, pero posteriormente se usó con otro tipo de administradores o ejecutivos. En una prueba de charola de pendientes, los candidatos a un puesto administrativo reciben una muestra de problemas del tipo que suele encontrarse en la lista de pendientes de un administrador (cartas, memorandos, notas, lineamientos, informes, mensajes telefónicos, correos electrónicos, faxes) y requieren de algún tipo de acción. A dichos candidatos se les solicita indicar qué medida debe tomarse en cada uno de los casos, y sus respuestas se evalúan de acuerdo con juicios de expertos sobre lo apropiado de las soluciones.

El método denominado *centros de evaluación*, introducido por la Compañía de Teléfonos y Telégrafos de Estados Unidos en la década de 1950, combina la técnica de charola de pendientes con otras tareas de simulación, tales como juegos de administración y ejercicios de resolución de problemas en grupo (como en la Prueba de Discusión en Grupo sin Líder). En este método también se utilizan entrevistas, pruebas psicológicas y otros procedimientos de evaluación. El centro de evaluación se ha empleado menos como una técnica de selección que como una forma de evaluar el personal de nivel gerencial para promoción y clasificación. Se instalan entre seis y doce candidatos en una ubicación específica, donde son observados y evaluados por otros ejecutivos entre sí durante varios días. Los principales criterios de examen son el grado de participación activa, las habilidades de organización y la habilidad para tomar decisiones.

Debido a que la técnica de charola de pendientes y otras tareas de simulación son realistas, podría parecer que resultan sumamente válidas. No obstante, los candidatos están conscientes de encontrarse "en un escenario" y pueden desempeñar cierto papel o comportarse de manera diferente a como lo harían en una situación administrativa real. Los gastos y las limitaciones de tiempo también impiden usar dichas técnicas de simulación para fines distintos a la evaluación de personal de gerencia de nivel bastante alto.

HABILIDADES SENSORIO-PERCEPTIVAS Y PSICOMOTRICES

Es importante, y en muchos casos está estipulado por el gobierno y otras organizaciones, que se evalúen periódicamente las habilidades sensorio-perceptivas y psicomotrices tanto de niños como de adultos. Excepto en casos de grandes deficiencias, no siempre es patente la presencia de un defecto en el funcionamiento físico. Dependiendo de que la desventaja pueda corregirse o compensarse y del grado en que afecte el desempeño en el trabajo, un aspirante a una institución educativa o empleo puede o no ser admitido o contratado. Sin embargo, la práctica actual, apoyada por el Acta de Estadounidenses con Discapacidad (ADA), favorece la contratación o admisión de personas con discapacidades y la adopción de medidas para reducir al mínimo los efectos debilitantes de sus desventajas.

Pruebas de visión y audición[1]

La agudeza tanto visual como auditiva puede verificarse mediante distintos tipos de pruebas, algunas de las cuales (tabla de Snellen, prueba de observación) son muy sencillas y otras (oftalmoscopio, audiómetro) mucho más complejas. Usualmente, un maestro o un asistente de

[1]En las páginas 61-75 de Fleishman y Reilly (1995) se encuentran ejemplos detallados de 12 pruebas de habilidades sensorio-perceptuales (visión cercana, visión lejana, discriminación visual del color, visión nocturna, visión periférica, percepción de profundidad, sensibilidad al brillo, sensibilidad auditiva, atención auditiva, ubicación del sonido, reconocimiento del habla, claridad del habla).

personal pueden aplicar pruebas sencillas de la vista y la audición, pero un examen más a fondo requiere los servicios de un optometrista, oftalmólogo o audiólogo profesional.

Un examen de la vista completo incluye pruebas de agudeza cercana y lejana para cada ojo y para ambos ojos juntos, el equilibrio muscular de los ojos a distancias cercanas y lejanas, percepción de profundidad y visión del color. Instrumentos tales como la Prueba de Visión B y L (de Bausch y Lomb) se han usado para revisiones de la vista en contextos industriales. Los resultados de la evaluación con estos instrumentos se evalúan en términos de diversas *familias de trabajos visuales*, dependiendo de qué habilidades visuales son esenciales para cada ocupación en particular. Una prueba de visión del color común consiste en una serie de cartas *seudo isocromáticas* que contienen un número o diseño formado por puntos coloreados contra un fondo de puntos contrastantes. La Prueba Dvorine de Visión del Color (The Psychological Corporation) es una prueba de este tipo de amplio uso.

Igual que una buena visión, un buen oído es importante en muchas ocupaciones, en particular en empleos como el de operador de sonares. La agudeza auditiva puede determinarse a grandes rasgos mediante una *prueba de reloj*,[2] pero una prueba profesional de audición implica el uso de un *audiómetro*. Los resultados de una prueba audiométrica se trazan en forma de gráfica (*audiograma*) donde se muestra la sensibilidad de cada oído a los tonos puros, los cuales cubren el rango de frecuencia de la audición humana. También puede determinarse la habilidad del individuo para ubicar la dirección de donde provienen los sonidos. Otra característica importante de la audición es la habilidad de discriminar entre estímulos de diferente tono o volumen.

Pruebas de habilidades psicomotrices

Las pruebas de habilidades psicomotrices figuraron entre las primeras medidas de habilidades especiales que se elaboraron. Muchas de las pruebas disponibles de este tipo se introdujeron en las décadas de 1920 y 1930 para predecir el desempeño en ciertos empleos u oficios calificados. Posteriormente, el Centro de Investigación de Capacitación y Personal de la Fuerza Aérea estadounidense realizó un amplio estudio de las habilidades psicomotrices que incluye el desempeño como piloto. De particular importancia en estos análisis era el desempeño en simuladores de vuelo como el Capacitador de Vínculo y la Prueba de Coordinación Compleja. En esta última el examinando usa un timón y tres controles similares a un bastón para ajustarse a un patrón de luces de estímulo que aparecen sobre un panel vertical para simular los movimientos de un aeroplano en vuelo.[3]

La velocidad, la fuerza y la agilidad, en conjunto, contribuyen al desempeño motriz efectivo. Las mediciones de estas características se usan ampliamente para seleccionar trabajadores en varios tipos de empleos y son pronosticadores válidos del desempeño en el trabajo físicamente demandante (vea Blakley, Quinones, Crawford y Jago, 1994; Hogan y Quigley, 1994). Además de las medidas de fuerza isométrica, están disponibles pruebas de precisión y estabilidad que implican varias manipulaciones con dedos, manos, brazos y piernas. Algunas de estas pruebas requieren de movimientos musculares pequeños, otras de movimientos grandes y otras más exigen tanto movimientos pequeños como grandes.

[2]En esta prueba, dependiendo de lo silenciosa que se encuentre la habitación del examen, una persona con oído normal debería ser capaz de oír el tic tac de un "reloj tamaño dólar" a una distancia de entre 75 cm y 1 m del oído.

[3]En las páginas 38-50 de Fleishman y Reilly (1995) se presentan ejemplos de pruebas de diez habilidades psicomotrices (precisión de control, coordinación multi-extremidades, orientación de la respuesta, control de tasas, tiempo de reacción, estabilidad brazo-mano, destreza manual, destreza de los dedos, velocidad muñeca-dedo, movimiento de velocidad de extremidades). En las páginas 51-60 de la misma fuente se incluyen ejemplos de pruebas de nueve habilidades físicas (fuerza estática, fuerza explosiva, fuerza dinámica, fuerza del tronco, flexibilidad de extensión, flexibilidad dinámica, coordinación corporal gruesa, equilibrio corporal grueso, vigor).

Para ilustrar las pruebas psicomotrices disponibles, a continuación se describirán algunas medidas seleccionadas de movimientos gruesos, finos o de una combinación de ambos tipos. La mayoría de estos instrumentos son apropiados tanto para adolescentes como para adultos, y se califican en términos de la cantidad de unidades de tarea terminadas en un tiempo específico o el lapso requerido para completar toda la tarea.

Movimientos manuales gruesos. Dos antiguas pruebas diseñadas para medir velocidad y precisión en los movimientos gruesos de dedos, manos y brazos son la Prueba Stromberg de Destreza (The Psychological Corporation) y la Prueba Minnesota de Índice de Manipulación (American Guidance Services). En la Prueba Stromberg de Destreza, se solicita al examinando colocar 54 discos de colores (rojo, amarillo, azul) del tamaño de una galleta en una secuencia preestablecida tan rápidamente como pueda (figura 10.1). Esta prueba se ha usado como medida de destreza manual en trabajadores de lavanderías, operadores de prensas cortadoras, moldeadores de máquinas, ensambladores y soldadores. La Prueba Minnesota de Índice de Manipulación consiste en un tablero de 60 orificios con bloques que son rojos por un lado y amarillos por el otro. La prueba se divide en cinco subpruebas, en las cuales los bloques se giran, mueven y colocan de ciertas maneras. En la parte de la Prueba de Colocación, por ejemplo, se colocan los bloques en los orificios del tablero; en la parte de Prueba de Giro, los bloques se giran y reemplazan en el tablero.

Movimientos manuales finos. Como representativas de las pruebas que requieren de manipulación de partes pequeñas figuran la prueba del Tablero de Clavijas Purdue (NCS London House) y la Prueba Crawford de Destreza con Partes Pequeñas (The Psychological Corporation). Se ha descubierto que las calificaciones de estas pruebas tienen correlaciones significativas con el desempeño en ocupaciones como mecánico de instrumentos, grabador, aguafuertista, ensamblador de electrónica de precisión y reparador de relojes.

El Tablero de Clavijas Purdue consiste en cinco tareas (mano derecha, mano izquierda, ambas manos, ensamblado de mano derecha más izquierda más ambas manos) para medir la destreza de mano-dedo-brazo requerida para ciertos tipos de trabajo manual. En la primera parte de la prueba, el examinando introduce alfileres en orificios, primero con la mano derecha, después con la izquierda, y al final con ambas manos. En la segunda parte, el examinando introduce

FIGURA 10.1 Prueba Stromberg de Destreza.

(Derechos reservados 1945, 1951, 1981 por The Psychological Corporation, una compañía de Harcourt Assessment Company. Reproducido con autorización. Todos los derechos reservados.)

un alfiler en un orificio, coloca una arandela y un aro sobre el alfiler, coloca otro alfiler en un orificio, y así sucesivamente (figura 10.2).

La Prueba Crawford de Destreza con Partes Pequeñas, que es una medida de la coordinación ojo-mano y de la destreza motriz fina, consta de dos partes. En la primera parte, el examinando utiliza pinzas para insertar alfileres en orificios y colocar aros sobre ellos. En la segunda parte coloca tornillos en orificios de rosca y los atornilla con un destornillador (figura 10.3).

Movimientos manuales gruesos y finos. La Prueba Bennett de Destreza Mano-Herramienta (The Psychological Corporation) es una prueba de habilidades psicomotrices que combina la destreza de los dedos con movimientos gruesos de los brazos. En esta prueba se solicita al examinando que primero saque 12 tuercas de 12 tornillos de tres tamaños diferentes montados sobre el lado izquierdo de un marco, y que después atornille de nuevo las tuercas y tornillos en el lado derecho del marco (figura 10.4). Las calificaciones consisten en el tiempo necesario para completar la tarea. En el manual se presentan las normas para administrar esta prueba a diversos grupos de aspirantes industriales.

Confiabilidad y validez de las pruebas de psicomotricidad

La confiabilidad de las pruebas de habilidades psicomotrices es inferior en promedio (.70 y .80) a la de otras pruebas de habilidades especiales. Una de las razones para que los coeficientes de confiabilidad resulten relativamente bajos en las pruebas de habilidades psicomotrices es que las calificaciones son sumamente susceptibles a la práctica (Fleishman, 1972).

En general, las pruebas de habilidades psicomotrices no han resultado muy útiles en la conserjería vocacional. Su validez suele ser inferior a la que tienen las pruebas de habilidades mecánicas y de trabajos de oficina. Las medidas de habilidades psicomotrices han sido más útiles para predecir el desempeño en programas de capacitación que para pronosticar la destreza en el empleo. También tienen mayor validez para prever el desempeño en trabajos repetitivos, tales como el ensamblado de rutina y la operación de máquinas, que en empleos complejos donde se involucran habilidades cognoscitivas y perceptuales de mayor nivel (Ghiselli, 1973).

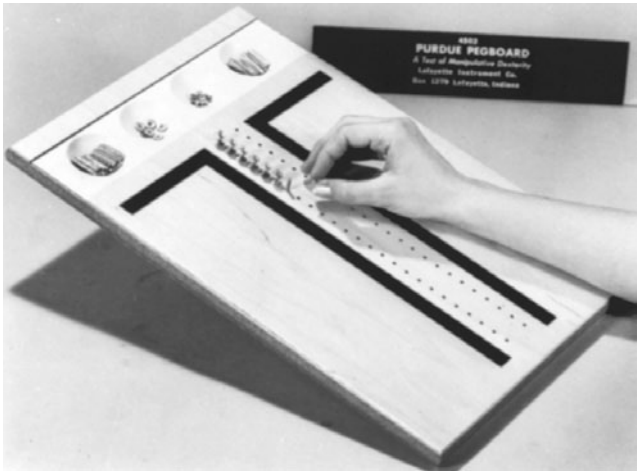


FIGURA 10.2 Tablero de Clavijas Purdue

(Cortesía de Lafayette Instrument Company.)

FIGURA 10.3 Prueba Crawford de Destreza con Partes Pequeñas, parte II.

(Derechos reservados 1946, 1956, 1981 por The Psychological Corporation, una Compañía de Evaluación de Harcourt. Reproducido con autorización. Todos los derechos reservados.)

FIGURA 10.4 Prueba Bennett de Destreza Mano-Herramienta.

(Derechos reservados 1969 por The Psychological Corporation, una Compañía de Evaluación de Harcourt. Reproducido con autorización. Todos los derechos reservados.)

HABILIDAD MECANICA

Se requiere un cierto nivel mínimo de habilidad psicomotriz para casi cualquier ocupación que involucre la operación de maquinaria; pero, más allá de ese nivel, la percepción espacial, el conocimiento mecánico y otras habilidades cognoscitivas son determinantes más importantes del desempeño. Uno de los primeros y más frecuentes tipos de habilidad especial que se mide es la habilidad mecánica. Hay algunas evidencias de un factor general débil de habilidad mecánica, pero las pruebas que se han diseñado para medirlo incluyen diversas habilidades perceptivo-motrices y cognoscitivas. Se trata de pruebas de habilidades psicomotrices, tales como la coordina-

ción muscular y de velocidad, la percepción de relaciones espaciales y la comprensión de relaciones mecánicas. Los componentes psicomotrices de diversas pruebas de habilidad mecánica, como las pruebas psicomotrices en general, tienen correlaciones bajas entre sí. Sin embargo, las correlaciones entre las calificaciones totales en diversas pruebas de habilidad mecánica a menudo son bastante considerables.

Un hallazgo interesante, aunque no es de sorprender, es la presencia de diferencias de género en las calificaciones de pruebas de habilidad mecánica. Es común que los varones obtengan calificaciones más elevadas en mediciones de comprensión espacial y mecánica, mientras que las mujeres logran calificaciones más altas en destreza manual fina y en ciertos aspectos de discriminación perceptual. Estas diferencias se tornan más pronunciadas en el bachillerato, y sin duda los factores sociales intervienen en su determinación.

Pruebas de relaciones espaciales

Un análisis intensivo de la habilidad mecánica, realizado por D. G. Paterson y sus colaboradores en la Universidad de Minnesota hacia finales de la década de 1920, condujo a la elaboración de tres pruebas: la Prueba Minnesota de Ensamblaje Mecánico, la Prueba Minnesota de Relaciones Espaciales, y el Tablero Minnesota de Formas de Papel (Paterson, Elliott, Anderson, Tooks y Heidbreder, 1930). La primera, una prueba de muestras de trabajo, requería que los examinados ensamblaran de nuevo un conjunto de objetos mecánicos desarmados. La tarea exigía destreza manual y percepción espacial, así como comprensión mecánica. El segundo y tercer instrumentos de esta serie eran pruebas de percepción espacial, habilidad considerada factor importante en los trabajos que involucraban tareas mecánicas. Como su nombre lo indica, la percepción espacial es la habilidad para visualizar objetos en tres dimensiones y manipularlos para producir una configuración particular.

Una descendiente de las pruebas anteriores fue la Prueba Minnesota de Relaciones Espaciales, edición revisada (American Guidance Service). Esta prueba, diseñada para edades de 16 años en adelante, evalúa la visualización espacial y la manipulación tridimensional de objetos. Consiste en cuatro tableros de formas (A, B, C, D) y dos series de bloques de formas geométricas. Una serie de bloques se ajusta en los huecos de los tableros A y B, y la segunda serie se ajusta a los huecos de los tableros C y D. La prueba empieza con los bloques dispersos fuera de los huecos, y se indica al examinando que tome los bloques y los coloque en los huecos correctos del tablero tan rápidamente como sea posible.

Otra descendiente de la Prueba Minnesota de Ensamblaje Mecánico es la Prueba Minnesota del Tablero de Formas de Papel, revisada (The Psychological Corporation). Esta adaptación a lápiz y papel de la Prueba Minnesota de Relaciones Espaciales fue diseñada para aplicarse desde el 9° hasta el 16° grados y en adultos. Consiste en 64 reactivos de opción múltiple, cada uno con un marco que muestra una figura geométrica dividida en varias partes y cinco marcos de respuestas que contienen una forma armada (figura 10.5). La tarea del examinando es seleccionar el marco de respuesta de los cinco que muestran cómo quedaría la figura al unir las partes entre sí. El Tablero Minnesota de Formas de Papel ha resultado útil para predecir grados en cursos de taller e ingeniería, así como para efectuar evaluaciones de supervisores y registros de producción en inspección, empaque, operación de máquinas y otras ocupaciones industriales. Las calificaciones de la prueba también se relacionan con el aprovechamiento en odontología y arte. Aunque se pretendía que la Prueba Minnesota del Tablero de Formas de Papel fuera una versión aplicada en forma más eficiente de la Prueba Minnesota de Relaciones Espaciales, la correlación entre calificaciones en las dos pruebas es considerablemente menor que el coeficiente de confiabilidad de pruebas paralelas del instrumento anterior.

Hay dos o más partes en la esquina superior izquierda para cada uno de los problemas presentados abajo. Elija entre las cinco figuras con las letras A, B, C, D, E, la que muestra cómo quedarían las partes de la esquina superior izquierda si se unieran entre sí. La respuesta correcta se muestra en el Problema 1.

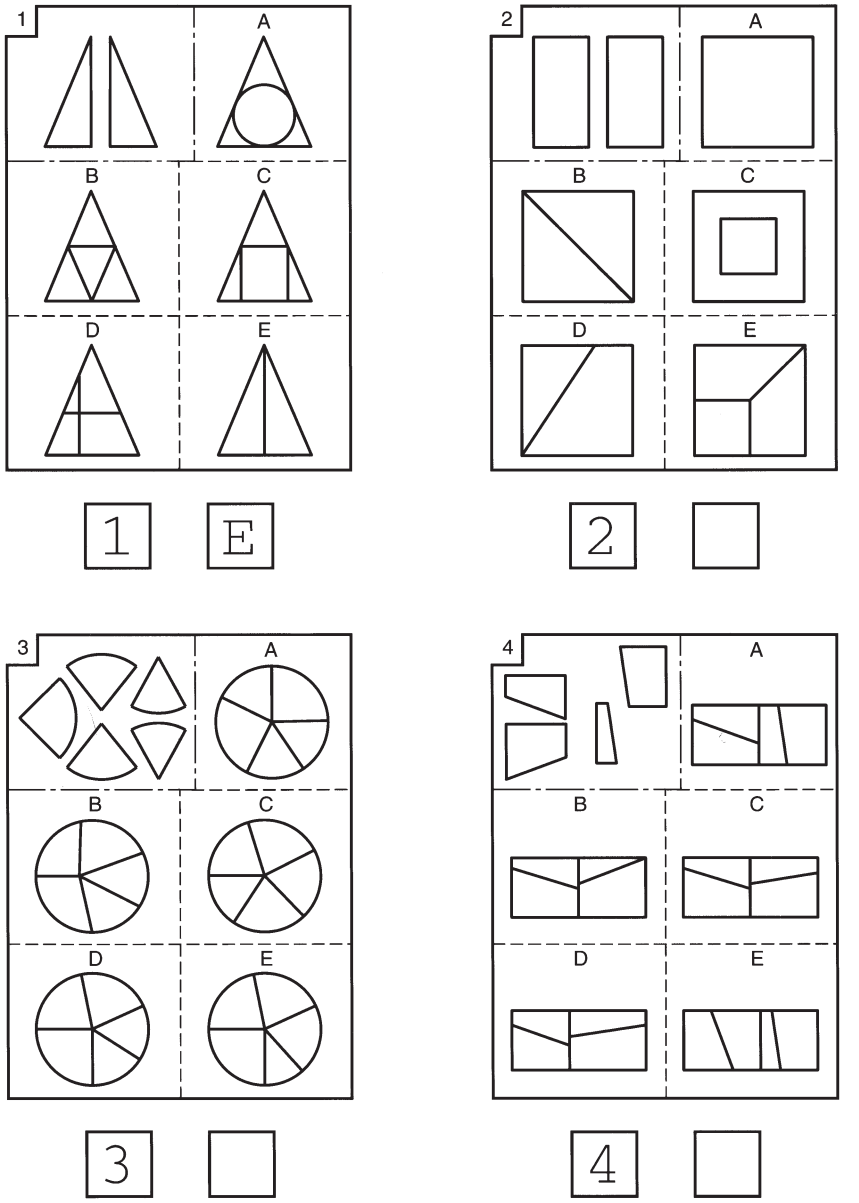


FIGURA 10.5 Muestra de reactivos de la Prueba Minnesota del Tablero de Formas de Papel, revisada.

(Derechos reservados © 1941, 1969 por The Psychological Corporation, una Compañía de Evaluación de Harcourt. Reproducido con autorización. Todos los derechos reservados. "RMPFBT" es una marca registrada propiedad de The Psychological Corporation e inscrita en Estados Unidos de Norteamérica y/u otras jurisdicciones.)

Otras medidas de habilidad mecánica en lápiz y papel

Ni la habilidad espacial ni la mecánica constan de un único factor. Por ejemplo, Carroll (1993) identificó cinco factores en pruebas de habilidad espacial: visualización, rotación acelerada, velocidad de cierre, flexibilidad de cierre y velocidad perceptiva. Los resultados del análisis factorial también indican que el desempeño en las pruebas de habilidad mecánica es una función de la habilidad espacial, la habilidad de razonamiento general, y la experiencia y conocimiento mecánicos (Alderton, 1994). Todos estos factores contribuyen a las calificaciones en pruebas de comprensión mecánica, las cuales se diseñan para evaluar la comprensión de los principios mecánicos involucrados en una gama de situaciones prácticas. Dos ejemplos de pruebas de este tipo son la Prueba de Conceptos Mecánicos (NCS London House) y la Prueba de Comprensión Mecánica Bennett (The Psychological Corporation). Las dos formas (S y T) de la Prueba de Comprensión Mecánica consisten en dibujos y preguntas sobre la operación de relaciones mecánicas y leyes físicas en situaciones prácticas (figura 10.6). La calificación y confiabilidad promedios de la prueba Bennett son menores para mujeres que para hombres, y se proporcionan normas separadas para cada sexo. Se encuentra evidencia de la validez de la prueba en sus modestas correlaciones con el desempeño en diversos trabajos mecánicos, técnicos y de manufactura.

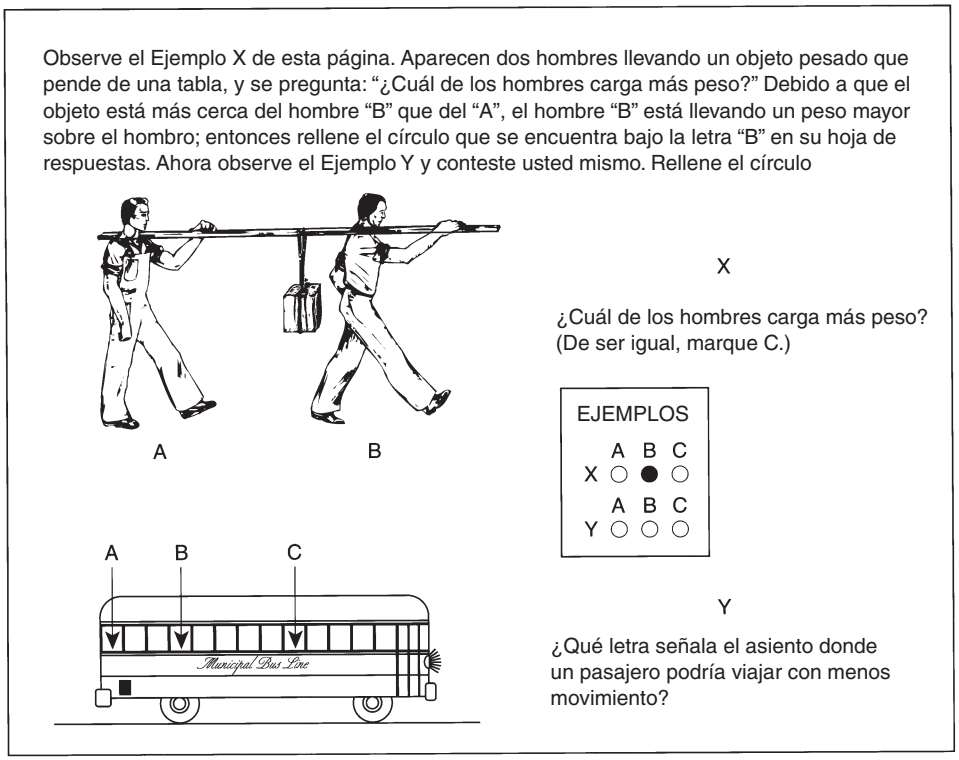


FIGURA 10.6 Ejemplo de reactivos de la Prueba de Comprensión Mecánica Bennett.

(Derechos reservados 1942, 1967-1970, 1980 por The Psychological Corporation, una Compañía de Evaluación de Harcourt. Todos los derechos reservados. "Bennett Mechanical Test" y "BMCT" son marcas registradas propiedad de The Psychological Corporation e inscritas en Estados Unidos y/u otras jurisdicciones.)

HABILIDADES PARA TRABAJOS DE OFICINA Y LAS RELACIONADAS CON LA COMPUTACIÓN

Al igual que muchas otras categorías de habilidad, la habilidad para desempeñar el trabajo de oficina no es un factor unitario distinto de la inteligencia general. La destreza manual y la velocidad para percibir semejanzas y diferencias son necesarias en el trabajo de oficina, pero las habilidades verbales y cuantitativas también son importantes. Por consiguiente, muchas pruebas de habilidades para el trabajo de oficina contienen reactivos similares a los que se encuentran en las pruebas de inteligencia general, así como reactivos para medir la velocidad y la precisión perceptual.

Además de las pruebas más generales de habilidad para el trabajo de oficina, se han diseñado algunos instrumentos para medir solamente la aptitud estenográfica. También hay disponibles pruebas sobre la capacidad para aprender las complejas tareas de oficina y de resolución de problemas de la programación y la operación de computadoras.

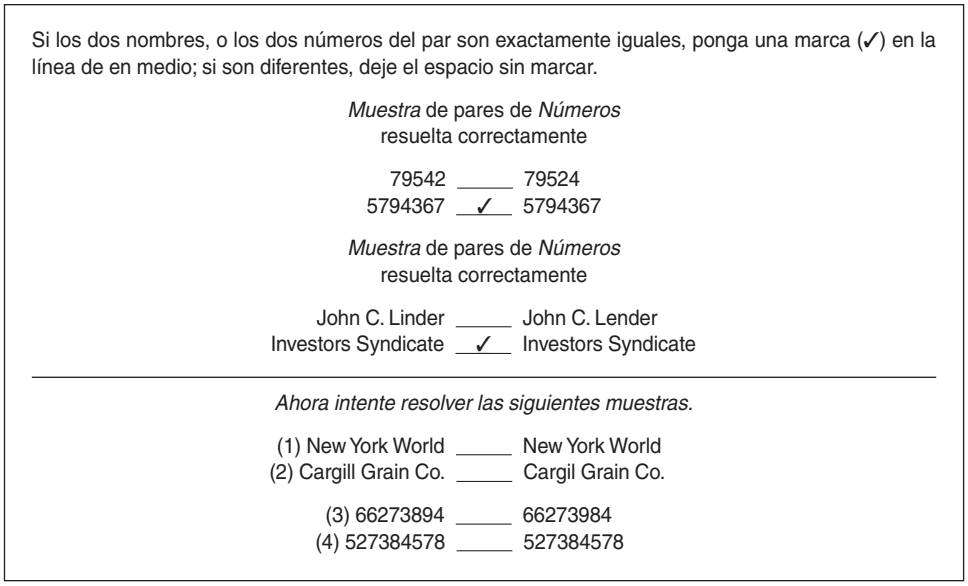
Pruebas representativas de la habilidad para el trabajo de oficina general

Las pruebas de habilidad para el trabajo de oficina que se distribuyen comercialmente varían en contenido, comprenden desde las tareas simples de marcar números y nombres en la Prueba Minnesota de Trabajo de Oficina, hasta las tareas combinadas perceptual-motrices y de inteligencia general de la reciente Batería de Habilidades para Trabajo de Oficina. La Prueba Minnesota de Trabajo de Oficina (The Psychological Corporation) fue diseñada para usarse al seleccionar empleados, inspectores y otros especialistas en ocupaciones que incluyen velocidad para percibir y manipular símbolos. Consiste en dos partes, Comparación de Números (ocho minutos) y Comparación de Nombres (siete minutos), en las que el examinando revisa 200 pares de números y 200 pares de nombres buscando errores y marca los pares idénticos (figura 10.7). Ambas partes se califican mediante la fórmula "aciertos menos errores". Las confiabilidades test-retest de las calificaciones están entre los .70 y 80. Las normas de rangos percentilares para estudiantes, por sexo y grado (7 a 12), y para grupos de trabajadores de oficina y solicitantes de empleo aparecen en el manual. Las calificaciones están moderadamente correlacionadas con las notas de maestros y supervisores sobre el trabajo de oficina.

En contraste con la Prueba Minnesota de Trabajo de Oficina, que sólo mide la velocidad y precisión de percepción, la Batería de Habilidades para Trabajo de Oficina (The Psychological Corporation) se compone de siete pruebas diseñadas para medir diversas habilidades del trabajo de oficina: Llenado (cinco minutos), Copiar información (cinco minutos), Comparación de información (cinco minutos), Uso de tablas (cinco minutos), Lectura de pruebas (cinco minutos), Habilidades matemáticas básicas (15 minutos) y Razonamiento numérico (20 minutos). Las normas de rangos percentilares basadas en varias poblaciones de empleados bien definidas están disponibles en las formas A y B de la prueba.

Habilidades relacionadas con la computadora

El rápido crecimiento de la industria de la computación durante las últimas décadas ha provocado cambios en la mayoría de los empleos de oficina y demanda de los programas de cómputo relacionados, es decir, la serie de enunciados lógicos que le dicen a la computadora qué hacer para lograr objetivos específicos. Aprender a programar computadoras y a usar los complejos programas que ya se han diseñado requiere de una combinación de las habilidades para el trabajo de oficina y para resolver problemas. Por lo tanto, es importante identificar a los individuos que poseen las aptitudes y habilidades necesarias para aprender cómo programar y manejar computadoras. Respondiendo a esta necesidad, los especialistas en mediciones han elaborado pruebas de



Si los dos nombres, o los dos números del par son exactamente iguales, ponga una marca (✓) en la línea de en medio; si son diferentes, deje el espacio sin marcar.

Muestra de pares de *Números*
resuelta correctamente

79542 _____ 79524
5794367 __✓__ 5794367

Muestra de pares de *Números*
resuelta correctamente

John C. Linder _____ John C. Lender
Investors Syndicate __✓__ Investors Syndicate

Ahora intente resolver las siguientes muestras.

(1) New York World _____ New York World
(2) Cargill Grain Co. _____ Cargil Grain Co.

(3) 66273894 _____ 66273984
(4) 527384578 _____ 527384578

FIGURA 10.7 Muestra de reactivos de la Prueba Minnesota de Trabajo de Oficina.

(Derechos reservados 1933. Renovados en 1961 por The Psychological Corporation, una Compañía de Evaluación de Harcourt. Reproducido con autorización. Todos los derechos reservados.)

aptitud para programar computadoras, uno de cuyos ejemplos es la Batería de Aptitud para Programador de Computadoras (CPAB) (NCS London House). Esta batería de prueba, que se diseñó para evaluar y seleccionar aspirantes a cursos de programación de computadoras, consiste en subpruebas de Significado Verbal, Razonamiento, Series de Letras, Habilidad para los Números, y Diagramación. El tiempo de evaluación es de 79 minutos para la versión Normal y de 55 minutos para la versión abreviada. Las pruebas se basaron en principiantes y programadores experimentados, y en analistas de sistemas, para desarrollar los reactivos, y en el manual se presentan las normas por rangos percentilares en las calificaciones totales y en las subpruebas. Los estudios sobre validación han demostrado que la CPAB puede predecir el éxito relevante en el trabajo para diversas compañías de producción y servicios.

También están disponibles pruebas que evalúan la capacidad para manejar computadoras. Un ejemplo es la Batería de Aptitud para Operador de Computadoras (COAB). Las tres subpruebas de esta batería de 45 minutos son: Reconocimiento de Secuencias ("habilidad para reconocer secuencias en forma rápida"), Marcar Formatos ("habilidad para percibir la adecuación de números y letras a un determinado formato") y Pensamiento Lógico ("habilidad de analizar problemas y visualizar soluciones de manera lógica"). El manual presenta normas de rangos percentilares, obtenidos en muestras relativamente pequeñas de operadores de computadoras con experiencia y aspirantes o aprendices sin experiencia, por subpruebas y por calificación total.

HABILIDADES ARTÍSTICAS Y MUSICALES

Las habilidades medidas por las pruebas de aptitud espaciales, mecánicas y de trabajo de oficina son importantes en ingeniería, mecánica, trabajo de oficina, odontología, y en cierta medida

en arte y música. Las pruebas de habilidades artísticas y musicales han sido objeto de décadas de investigación, y algunas de ellas tienen al menos una validez modesta. No obstante, la investigación dedicada a la evaluación de estas habilidades ya no se lleva a cabo con la energía con que alguna vez se realizó (Carson, 1998).

Pruebas de aptitud artística

Como dice la frase: "la belleza está en los ojos de quien la contempla", el juez último sobre el mérito artístico es el observador. Debido a que el gusto en el arte varía mucho de una persona a otra, de una cultura a otra y de generación a generación, no es sorprendente que los criterios sobre la habilidad artística resulten difíciles de especificar. Sin importar los múltiples problemas que se presentan al tratar de definir criterios confiables y elaborar instrumentos para predecirlos, se han publicado varias pruebas de habilidad para el arte visual y la aptitud musical. Sin embargo, muchas de estas pruebas son obsoletas y ya no están disponibles comercialmente.

Hace algunos años los investigadores de la Universidad de Minnesota encontraron una correlación positiva entre las calificaciones en pruebas de percepción espacial, tales como la Prueba Minnesota del Tablero de Formas de Papel, y la aptitud artística (Paterson *et al.*, 1930). Desde luego que la habilidad espacial no es el único factor que cuenta en la aptitud artística; el juicio, la destreza manual, la imaginación creativa y otros factores también intervienen. Asimismo, una persona que puede reconocer el buen arte no necesariamente es capaz de producirlo. Por ello, es importante distinguir entre medidas de apreciación estética (juicio y percepción) y medidas de muestras de trabajo de habilidad productiva en arte. Como ejemplo de pruebas de juicio y percepción del arte están la Prueba Meier de Juicio Artístico, la Prueba Meier de Percepción Estética (Meier, 1942), y la Prueba Graves de Juicio de Diseño (Graves, 1948). A diferencia de la Prueba Meier de Juicio Artístico, que utiliza obras de arte famosas como material de prueba, la de Graves emplea obras abstractas de dos y tres dimensiones para revelar los juicios artísticos. Ejemplo de una prueba de desempeño en arte es el Inventario Horn de Aptitudes Artísticas (Horn y Smith, 1945), donde se requiere que el examinando esboce objetos comunes y figuras geométricas y trace conjuntos de líneas básicas en marcos rectangulares.

Pruebas de aptitud musical

No está clara la importancia relativa de la habilidad innata, la motivación, la instrucción y la práctica para la determinación del talento musical. Hay ciertas muestras de que existe un factor general débil de la aptitud musical, pero la mayoría de las investigaciones han demostrado que son varias las habilidades que contribuyen al logro musical. Uno de dichos factores es la habilidad de discriminar entre diferentes tonos, el *tono perfecto* que supuestamente ha caracterizado a muchos músicos famosos. Como lo revela la investigación que ha usado técnicas de sondeo cerebral, los factores neuropsicológicos son importantes para determinar el tono perfecto. Por ejemplo, Schlaug, Jaencke, Huang y Steinmetz (1995) descubrieron que los músicos con tono perfecto tenían una marcada asimetría del plano temporal izquierdo que los no músicos o los músicos sin tono perfecto. En una investigación relacionada, Schlaug *et al* (1995) encontró que la mitad anterior del cuerpo calloso era significativamente mayor en los músicos profesionales que en los no músicos.

La prueba de aptitud musical más antigua, denominada Medidas Seashore de los Talentos Musicales, fue producto de la investigación pionera de Carl Seashore y sus colegas en la Universidad de Iowa durante las décadas de 1920 y 1930 (Seashore, 1939). En contraste con las pruebas de aptitud musical que se desarrollaron más tarde, los materiales de estímulo de las pruebas

de Seashore consistían en un conjunto de tonos o notas musicales, más que en selecciones musicales significativas. Este método analítico, atomista de medir la aptitud musical fue seriamente criticado y, en consecuencia, se desarrollaron varias pruebas con un contenido más complejo. Entre éstas están las medidas colectivas como la Prueba Drake de Aptitud Musical (Drake, 1954) y el Perfil de Aptitud Musical (por E. E. Gordon, de GIA Publications).

El Perfil de Aptitud Musical (MAP) es una prueba grabada consistente en 250 selecciones breves originales para violín y violonchelo tocadas por músicos profesionales. No se requiere de ningún conocimiento previo sobre hechos musicales o históricos. El MAP consta de tres pruebas que miden siete componentes: Imaginación Tonal (melodía y armonía), Imaginación Rítmica (ritmo y métrica), y Sensibilidad Musical (fraseo, equilibrio y estilo). Según su revisión en 1995, la prueba toma aproximadamente tres horas y media en aplicarse y puede calificarse manualmente.

Otras pruebas de música diseñadas por E. E. Gordon y publicadas por GIA Publications incluyen:

Medidas Avanzadas de Audición de Música. Prueba de aptitud musical para estudiantes universitarios, dura 20 minutos y proporciona calificaciones sobre tono, ritmo y compuestas.

Prueba de Preferencia de Timbre de Instrumento. Ayuda a los estudiantes a partir de los nueve años de edad a seleccionar un instrumento de aliento apropiado, de metal o de madera, para aprender a tocar.

Registro de Prontitud para la Improvisación Armónica y Registro de Prontitud para la Improvisación. Diseñada para funcionar como una ayuda objetiva para maestros (desde el tercer grado hasta escuelas de especialización musical) y ayudar a los alumnos a improvisar música.

Prueba Iowa de Alfabetismo Musical. Una prueba de aprovechamiento musical estandarizada nacionalmente, para grados del cuarto al duodécimo, diseñada para evaluar el progreso, diagnosticar cualidades y deficiencias, y comparar la posición relativa de los alumnos en el aprovechamiento musical.

BATERÍAS DE PRUEBAS DE APTITUDES MÚLTIPLES

En la conserjería vocacional, así como en la clasificación y colocación de empleos, a menudo resulta útil evaluar las habilidades y el conocimiento en varias áreas. Un consejero puede decidir aplicar una serie de pruebas individuales de habilidades, pero éste puede no ser el procedimiento más eficaz. Además, es probable que un conjunto de pruebas separadas se haya estandarizado en tantos grupos de personas distintos como pruebas existan. Debido a que los grupos de norma pueden variar en forma significativa, es difícil establecer una comparación relevante entre la calificación de una persona en una prueba y sus calificaciones en otras pruebas.

Las pruebas separadas de capacidades especiales ciertamente tienen un lugar, sobre todo en la selección y el sondeo de personal, pero son menos útiles en la asesoría y el diagnóstico vocacional. El énfasis en los procedimientos de selección de personal ha cambiado un poco durante las últimas décadas de seleccionar sólo la "crema y nata" a clasificar y colocar a los trabajadores en los empleos más adecuados a sus habilidades y necesidades. Por consiguiente, la aplicación de las pruebas comprendidas en una batería de múltiples habilidades, la cual está diseñada

para asignar a las personas con patrones de habilidades particulares a empleos específicos, se considera más eficaz que aplicar una serie de pruebas no relacionadas diseñadas para seleccionar sólo a los mejores y descartar a todos los demás. A diferencia de las pruebas únicas de habilidades especiales, que pueden ser del tipo de lápiz y papel o de ejecución, una batería de habilidades múltiples típica no requiere más dispositivos que lápiz y papel y puede administrarse simultáneamente a un grupo grande de estudiantes, aspirantes a un empleo, reclutas militares y otros grupos.

Debido a que las habilidades cognoscitivas son menos específicas durante los años de la escuela elemental, en general, antes de cursar la secundaria no se recomienda aplicar una batería cara, que tome mucho tiempo, de habilidades múltiples. Durante la secundaria, al ir diferenciándose sus habilidades cognoscitivas con la madurez y la experiencia, los alumnos empiezan a investigar y hacer planes sobre sus futuras carreras, así como a decidir los cursos académicos que tomarán. Para ayudarlos en estos esfuerzos, muchos sistemas escolares aplican una batería de habilidades múltiples en el octavo o noveno grado de bachillerato. La información proporcionada por la batería de pruebas puede aumentar la conciencia de los alumnos sobre sus cualidades y deficiencias y, por lo tanto, guiarlos para tomar decisiones laborales y educativas.

Más que aplicar una larga serie de pruebas de habilidades especiales o una batería de habilidades múltiples, un consejero vocacional puede decidir usar una prueba de inteligencia general y una o más pruebas de habilidades especiales. Ciertamente, nada tiene de malo esta estrategia, porque las habilidades verbales y cuantitativas medidas por las pruebas de inteligencia son importantes en un amplio espectro de posiciones académicas y vocacionales. Además de evaluar varias habilidades especiales, muchas baterías de habilidades contienen una prueba de inteligencia general. Esto proporciona las ventajas combinadas de una aplicación más eficiente y de normas comparables en todas las pruebas.

Diferencias de calificaciones e interpretación de perfiles

Los procedimientos estadísticos del análisis factorial (vea apéndice A) se han usado en la elaboración de varias baterías de pruebas de habilidades. Incluso en baterías no desarrolladas por métodos de análisis factorial, los resultados de estudios que emplean estos métodos usualmente se han considerado para elaborar los reactivos y definir las variables que habrán de medirse.

Los reactivos de las Pruebas de Aptitudes Diferenciales, una de las más populares baterías de pruebas usadas en la asesoría académica en el nivel de bachillerato, se seleccionaron para que tuvieran correlaciones elevadas con otros reactivos en la misma subprueba, pero correlaciones bajas con reactivos de otras subpruebas. El resultado final fue un conjunto de subpruebas internamente consistentes con bajas correlaciones entre sí. Era importante que las correlaciones entre subpruebas fueran bajas; de lo contrario, el traslape entre las habilidades medidas por distintas subpruebas sería demasiado grande como para obtener una interpretación diferencial de las calificaciones de la subprueba.

Confiabilidad y error estándar de las diferencias entre calificaciones. Las magnitudes de las correlaciones entre distintas subpruebas en la misma batería a menudo son notables, y el hecho de que las subpruebas son bastante breves ocasiona que su confiabilidad sea muy baja. No sólo la confiabilidad de las diferencias entre calificaciones en dos pruebas varía directamente con la confiabilidad de las pruebas, también varía inversamente con la correlación entre las pruebas. La correlación considerable entre dos subpruebas dadas, combinada con su baja confiabilidad, origina que la confiabilidad de las diferencias entre calificaciones de las subpruebas sea baja.

La siguiente es una fórmula de la confiabilidad de las diferencias (r_{dd}) entre las calificaciones de las mismas personas en dos pruebas o subpruebas con iguales varianzas:

$$r_{dd} = \frac{r_{11} + r_{22} - 2r_{12}}{2(1 - r_{12})}, \quad \textbf{(10.1)}$$

donde r_{11} es la confiabilidad del primer conjunto de calificaciones, r_{22} la confiabilidad del segundo conjunto de calificaciones, y r_{12} la correlación entre estos dos conjuntos. Por ejemplo, supóngase que la confiabilidad de calificaciones anteriores a la prueba es $r_{11} = .90$, la confiabilidad de las calificaciones posteriores a la prueba es de $r_{22} = .80$, y la correlación entre calificaciones anteriores y posteriores a la prueba es $r_{12} = .70$. Entonces, la confiabilidad de la diferencia entre calificaciones anteriores y posteriores a la prueba es $[.90 + .80 - 2(.70)]/[2(1 - .70)] = .50$.

Cuando la confiabilidad de las diferencias entre calificaciones en dos subpruebas es baja, la diferencia entre las calificaciones de una persona en la subpruebas debe ser bastante grande para que pueda resultar significativa. Con el propósito de ilustrar este principio, supongamos que la confiabilidad de las calificaciones *T* en la prueba de relaciones espaciales de una batería de habilidades es .85, y la confiabilidad de las calificaciones *T* en la prueba de aptitud mecánica de la misma batería es .90. Un valor aproximado del error estándar de las diferencias (s_{ed}) entre calificaciones en dos pruebas con desviaciones estándar iguales puede calcularse mediante:

$$s_{ed} = s\sqrt{2 - r_{11} - r_{22}}, \quad \textbf{(10.2)}$$

donde r_{11} y r_{22} representan la mejor confiabilidad de test-retest de las dos pruebas, y s es la desviación estándar de las calificaciones de cada prueba. Recordando que la desviación estándar de las calificaciones *T* es igual a 10, cuando las calificaciones de ambas pruebas se expresan como calificaciones *T*, la fórmula 10.2 se convierte en:

$$s_{ed} = 10\sqrt{2 - .85 - .90} = 5.$$

En consecuencia, para estar bastante seguros (digamos con una probabilidad de .95) de que la diferencia entre las calificaciones de una persona en estas dos pruebas no se debe al azar, tal diferencia debe ser de al menos $1.96 \times 5 = 9.8$ unidades de calificaciones *T*.

Perfil de calificaciones. El proceso de interpretar la calificación de una persona en una batería de aptitudes múltiples constituida por varias pruebas estandarizadas en los mismos grupos de norma o en grupos equivalentes empieza con la construcción de un perfil de calificación. Un *perfil de calificación*, que es una gráfica de líneas o de barras de las calificaciones obtenidas en distintas pruebas, proporciona una imagen de las ventajas y deficiencias de una persona en varias áreas de aptitud. A partir de las normas es posible construir un perfil de las calificaciones de la persona en diversas pruebas para su uso en asesoría académica o vocacional. Más que trazar las calificaciones de una persona como puntos específicos en una gráfica, éstas pueden representarse como una serie de barras horizontales o verticales que van de uno o dos errores de medición estándar a cada lado de la calificación (vea la figura 5.2, en la página 92). Entonces, si las barras verticales de las dos pruebas no se traslapan, la diferencia entre las calificaciones de la persona en estas dos pruebas puede interpretarse como significativa.

Con el fin de apoyar la conserjería vocacional y la selección y colocación ocupacional, puede ser útil comparar el perfil de las calificaciones de una persona en una batería de habilidades múltiples con el perfil de las calificaciones promedio de la gente ubicada en ocupaciones seleccionadas para tal efecto. Aunque los trabajadores de una misma ocupación difieren en cierta

medida en cuanto a sus patrones de habilidad, ciertas familias de empleos parecen requerir un conjunto particular de habilidades. Los perfiles similares en una batería de aptitudes múltiples indican patrones similares de habilidades.

Pruebas de aptitud diferencial

Cierto número de baterías de pruebas de aptitud han sido diseñadas para, y estandarizadas principalmente en, situaciones escolares y se han usado para pronosticar el aprovechamiento académico. Una importante batería de este tipo se conoce como Pruebas de Aptitud Diferencial (DAT) (The Psychological Corporation). Las DAT se han usado sobre todo para consejería educativa y vocacional en estudiantes de bachillerato, pero también se ha empleado en educación básica para adultos, programas universitarios, vocacional-técnicos y correccionales. Hay dos niveles de la última (quinta) edición de las DAT: el Nivel 1 para grados del 7° al 9°, y el Nivel 2 para grados del 10° al 12°.

Las DAT comprenden ocho pruebas: Razonamiento Verbal, Razonamiento Numérico, Razonamiento Abstracto, Velocidad y Exactitud Perceptual, Razonamiento Mecánico, Relaciones Espaciales, Ortografía y Uso del Lenguaje. El tiempo para resolver toda la batería es de 156 minutos, pero una Batería Parcial de las DAT consistente en Razonamiento Verbal y Razonamiento Numérico requiere sólo de 90 minutos de trabajo. Además, se ha adaptado una edición computarizada de toda la batería que sólo toma 90 minutos en promedio para resolverse. Usando la teoría de respuesta al item, la versión adaptada presenta un subconjunto de reactivos de prueba que son los más adecuados para la persona que se somete a la prueba.

La quinta edición de las DAT fue estandarizada sobre una muestra nacional representativa de los estudiantes estadounidenses de bachillerato, se estratificó de acuerdo con el tamaño del distrito escolar, la región geográfica, el estatus socioeconómico de la comunidad y el tipo de escuela (pública o privada). Hay considerables diferencias de género en las calificaciones de las DAT: las mujeres tienen calificaciones más elevadas en Uso del Lenguaje y en Velocidad y Exactitud Perceptual, mientras que los hombres obtienen mayores puntuaciones en Razonamiento Mecánico y Relaciones Espaciales. Debido a estas diferencias de género, las normas por rango percentilar, estanina, y de la calificación escalada se presentan por separado para hombres y mujeres, así como para ambos sexos combinados.

Los datos estadísticos en la quinta edición del manual para el DAT, así como para la prueba misma, son bastante añejos, lo que debe tomarse en cuenta al usar esta batería de pruebas. Los coeficientes de la consistencia interna de las ocho pruebas varían de .82 a .95, y los coeficientes de formas paralelas están entre .73 y .90. Las correlaciones entre las pruebas van desde casi cero entre la Velocidad y Exactitud Perceptual y otras pruebas de la batería, hasta .70 entre las pruebas de Razonamiento y Uso del Lenguaje. Los extensos datos presentados en el manual indican que las escalas de las DAT, y especialmente el Razonamiento Verbal más el Razonamiento Numérico en conjunto, son válidas para pronosticar los grados de bachillerato y universitario. La batería de las DAT es útil para predecir el nivel del empleo dentro de las ocupaciones, pero las normas para varias ocupaciones son limitadas. En consecuencia, como pronóstico diferencial del éxito vocacional, las escalas de las DAT deben usarse con precaución.

Batería Multidimensional de Aptitudes, II

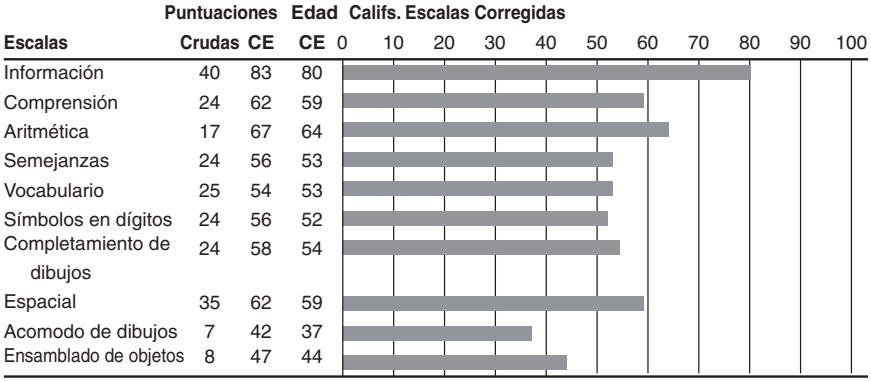
La Batería Multidimensional de Aptitudes, II (MAB-II) es una adaptación administrada de manera colectiva de la Escala de Inteligencia para Adultos de Wechsler, revisada. Como la WAIS-R, la MAB-II consiste en dos escalas (Verbal y de Desempeño) con cinco subpruebas cada una. Las cinco subpruebas Verbales son: Información, Comprensión, Aritmética, Semejanzas y Vo-

cabulario. Las cinco subpruebas de Desempeño son: Símbolos en Dígitos, Completamiento de Dibujos, Espacial, Acomodamiento de Imágenes y Ensamblado de Objetos. El tiempo límite para cada subprueba es de siete minutos, de modo que toda la batería puede terminarse en menos de hora y media. Es posible obtener calificaciones estándar y CI a partir de las baterías Verbal, de Ejecución y de Escala Completa, así como calificaciones escaladas de subpruebas, y un informe detallado de las calificaciones y su interpretación, a través del servicio de calificaciones por computadora de los Sistemas de Asesoría Sigma. En la figura 10.8 se muestra un ejemplo de perfil de calificaciones de la MAB-II.

El manual MAB-II (Jackson, 1998) registra la confiabilidad test-retest durante un periodo de 45 días como de .95 para la calificación Verbal, de .96 para la de Ejecución , y de .97 para la de Escala Completa. En un estudio con 500 personas de entre 16 y 20 años de edad, los coeficientes de consistencia interna para los CI Verbal, de Ejecución y de Escala Completa estaban entre los altos niveles de .90. Las correlaciones entre las calificaciones MAB y los CI de la WAIS-R en una muestra de 145 adultos fueron de .94 para la prueba Verbal, de .79 en Ejecución y de .91 para la Escala Completa de calificaciones WAIS-R . Los resultados de los análisis factoriales de las calificaciones de las subpruebas indican que, al igual que la Wais-R, la batería MAB-II mide un factor de inteligencia general así como factores de ejecución y verbal por separado.

Batería de Pruebas de Aptitud General

Se han diseñado varias baterías de pruebas de habilidades múltiples específicamente para la selección y colocación de personal en el medio empresarial e industrial. Entre éstas se encuentra



Las puntuaciones crudas para cada prueba indican la cantidad de preguntas que el examinado respondió en forma correcta. El primer conjunto de Calificaciones Escaladas (CE) no está basado en la edad y se usa para calcular las calificaciones de CI Verbal, de Ejecución y de Escala Completa. Las Calificaciones Escaladas (CE) corregidas por edad y la barra gráfica correspondiente comparan los resultados de los examinados con los de las personas del mismo grupo de edad.

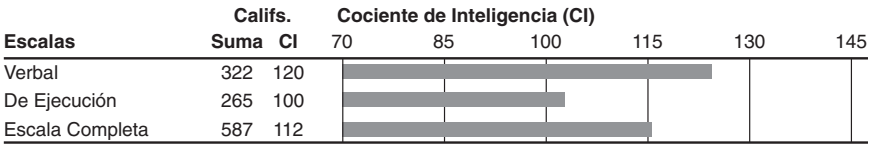


FIGURA 10.8 Perfil de Calificaciones en la Batería Multidimensional de Aptitudes, II.

(Reproducida con autorización de Sigma Assessment Systems, Inc., P.O. Box 610984, Port Huron, MI 48061-0984. Tel. (800) 265-1285.)

la Batería de Pruebas de Aptitud General, las Pruebas Flanagan de Clasificación de Aptitud, las Pruebas Industriales Flanagan y el Estudio de Habilidades para Empleados. Una de las más antiguas baterías de prueba orientadas hacia la industria fue diseñada en la década de 1930 por el personal del Instituto de Investigación para la Estabilización del Empleo de Minnesota (MESRI). La batería MESRI contenía pruebas de inteligencia general, así como para medir la habilidad numérica, perceptual, mecánica y psicomotriz. Los perfiles de las calificaciones promedio de estas pruebas obtenidas por empleados de oficina, trabajadores mecánicos, vendedores y muchos otros grupos ocupacionales, se establecieron como un conjunto de *patrones de habilidad ocupacional* (*OAP*) con los cuales podía compararse el desempeño individual.

El enfoque OAP de la batería MESRI se conservó al desarrollarse la Batería de Pruebas de Aptitud General (GATB) del Servicio de Empleo de Estados Unidos. La GATB, diseñada con base en el análisis del empleo y en un análisis factorial de 59 pruebas, se compone de ocho pruebas de lápiz y papel y cuatro pruebas con aparatos. Estas 12 pruebas en conjunto producen calificaciones sobre nueve habilidades principales requeridas para el éxito laboral: Habilidad de Aprendizaje General (G), Habilidad Verbal (V), Habilidad Numérica (N), Habilidad espacial (S), Percepción de Formas (P), Percepción del Trabajo de Oficina (Q), Coordinación Motriz (K), Destreza de Dedos (F) y Destreza Manual (M). Las puntuaciones crudas en estas nueve variables se convierten a rangos percentilares o calificaciones estándar con una media de 100 y una desviación estándar de 20. Estas tres calificaciones compuestas se obtienen de combinaciones apropiadas de las puntuaciones alcanzadas en los nueve factores: Cognoscitivo = G + V + N, Perceptual = S + P + Q y Psicomotriz = K + F + M.

Las calificaciones estándar de una persona en las variables GATB pueden compararse con las de los aproximadamente 36 patrones de habilidad ocupacional (OAP) determinados a partir de un análisis de las calificaciones de personas en más de 800 empleos. Un OAP consiste en un conjunto de calificaciones de GATB mínimas consideradas esenciales para el desempeño eficaz en determinada ocupación.

Toda la batería GATB tarda $2\frac{1}{2}$ horas en administrarse y es adecuada para los últimos grados de bachillerato (en general, el 12° grado) y adultos. Los coeficientes de confiabilidad test-retest y de formas paralelas para las pruebas separadas van de .80 a .90, con un error estándar de medición de aproximadamente 7 puntos para las calificaciones estándar. La validez de las nueve pruebas de habilidad y los 36 OAP para predecir los criterios de éxito ocupacional y académico está entre .00 y .90.

La GATB ha sido una de las herramientas más usadas en consejería vocacional y colocación laboral para estudiantes de los grados 9° al 12° y adultos, y probablemente es la batería de pruebas más adecuada para tal propósito. Debido a la supuesta injusticia de la GATB para grupos minoritarios, en 1981 se aplicó un sistema de *normas por raza* para las calificaciones como parte del programa de acción afirmativa del Departamento del Trabajo de Estados Unidos. Esta medida consistía en usar normas de rangos percentilares separadas para blancos, negros y latinos, y en registrar sólo los rangos percentilares de candidatos dentro del grupo. Sin embargo, los críticos consideraron esta práctica como *discriminación invertida* y en 1990 el uso de la GATB fue suspendido por el Departamento de Justicia mientras podían arreglarse los problemas concernientes a la justicia y la discriminación invertida. El Congreso de Estados Unidos también incluyó al lenguaje en el Acta de Derechos Civiles de 1991 para efectos de que ya no se aplicaran normas y ajustes a las calificaciones para diferenciar grupos.[4] Posteriormente el Servicio de Em-

[4]Al parecer está en desacuerdo con la disposición del Acta de Derechos Civiles de 1991, que evitaba los ajustes de calificaciones para las diferencias de grupos, el requisito del Acta de Estadounidenses con Discapacidad, de 1990, de que los jefes proporcionen colocación en la evaluación para individuos con discapacidades sensoriales, manuales o del habla. Sin embargo, como lo señala Tenopyr (1996), eso es un desacuerdo sobre una medida de política pública más que un conflicto científico o psicométrico.

pleo de Estados Unidos reanudó el uso de la GATB, pero los informes para los empleadores ya no están regulados por la raza. Más bien, las puntuaciones crudas para varias de las pruebas que conforman la batería se convierten a calificaciones estándar con base en normas combinadas para todos los grupos raciales.

Batería de Aptitud Vocacional de las Fuerzas Armadas

A partir de los Exámenes Army Alfa y Army Beta en la Primera Guerra Mundial, a lo largo de los años se han utilizado diversas pruebas para seleccionar y clasificar al personal de las Fuerzas Armadas de Estados Unidos. La Prueba de Clasificación General del Ejército (AGCT) y la Prueba de Clasificación General de la Marina (NGCT) se aplicaron a millones de reclutas militares durante y después de la Segunda Guerra Mundial para clasificarlos en trabajos capacitados y no capacitados, para seleccionar a quienes podrían beneficiarse de una mayor capacitación y rechazar a aquellos que, por una habilidad mental baja, se consideraran inadecuados para el servicio militar (vea Harrell, 1992). Algunos años después de la Segunda Guerra Mundial, la Prueba de Capacitación de las Fuerzas Armadas (AFQT) reemplazó a la AGCT y a la NGCT.

Durante la década de 1970, la Batería de Aptitud Vocacional de las Fuerzas Armadas se convirtió en la prueba uniforme de selección y clasificación de los servicios armados en conjunto. La forma actual de esta batería (ASVAB, 18/19), que es la prueba de habilidades múltiples más aplicada en Estados Unidos, consiste en diez pruebas:

Ciencia General (*GS*): 25 reactivos que miden el conocimiento en ciencias físicas y biológicas.

Razonamiento Aritmético (*AR*): 30 reactivos que miden la capacidad de resolver problemas verbales de aritmética.

Conocimiento de Palabras (*WK*): 35 reactivos para medir la habilidad de seleccionar el significado correcto de las palabras presentadas en contexto e identificar el mejor sinónimo de una palabra determinada.

Comprensión de Párrafos (*PC*): 15 reactivos que miden la habilidad para obtener información a partir de textos escritos.

Operaciones Numéricas (*NO*): 50 reactivos que miden la habilidad de realizar cálculos aritméticos.

Velocidad de Codificación (*CS*): 84 reactivos que miden la habilidad para usar una clave y asignar códigos de números a palabras.

Información sobre Autos y Talleres (*AS*): 25 reactivos que miden el conocimiento sobre automóviles, herramientas y terminología y prácticas de talleres.

Conocimiento Matemático (*MK*): 25 reactivos que miden el conocimiento de principios matemáticos a nivel de bachillerato.

Comprensión Mecánica (*MC*): 25 reactivos que miden el conocimiento de principios mecánicos y físicos y la habilidad para visualizar cómo funcionan las herramientas de trabajo ilustradas.

Información sobre Electrónica (*EI*): 20 reactivos para medir el conocimiento sobre electricidad y electrónica.

El tiempo de administración para las pruebas de ASVAB varía desde 3 minutos para las Operaciones Numéricas hasta 36 minutos para el Razonamiento Aritmético, con un total de 144

minutos para completar las diez pruebas. Se registran calificaciones *T* estándar y bandas de calificaciones de rangos percentilares para cada una de las pruebas y tres calificaciones compuestas: Habilidad Verbal (VA) = WK + PC, Habilidad Matemática (MA) = AR + MK, y Habilidad Académica (AA) = VA + MA. Las calificaciones de cuatro combinaciones ocupacionales también pueden calcularse como combinaciones adecuadas de calificaciones en las diez pruebas básicas: Mecánica y Oficios (MC); Negocios y Trabajo de Oficina (BC); Electrónica y Electricidad (EE), Salud, Social y Tecnología (HST).

Como se ilustra en la figura 10.9, el desempeño de una persona en la ASVAB puede representarse mediante una serie de bandas de calificaciones de rangos percentilares que indican los rangos dentro de los cuales es probable que caigan las verdaderas calificaciones de las pruebas de una persona. Además de trazar las bandas de calificaciones de rangos percentilares de mismo grado-mismo sexo, pueden trazarse las calificaciones *T* de mismo grado-mismo sexo y mismo grado-sexo opuesto. Dos datos adicionales que se incluyen en la hoja del perfil son los códigos primario y secundario de la ASVAB de la persona y la Calificación de las Carreras Militares. Los códigos se interpretan mediante un OCCU-FIND especial en un cuaderno diseñado para ayudar a los examinados a identificar las ocupaciones militares con las que concuerdan más sus calificaciones. La Calificación de las Carreras Militares, la cual se usa en conjunto con las gráficas de un folleto de *Carreras Militares* que se proporciona al examinado, colabora en el proceso de eva-

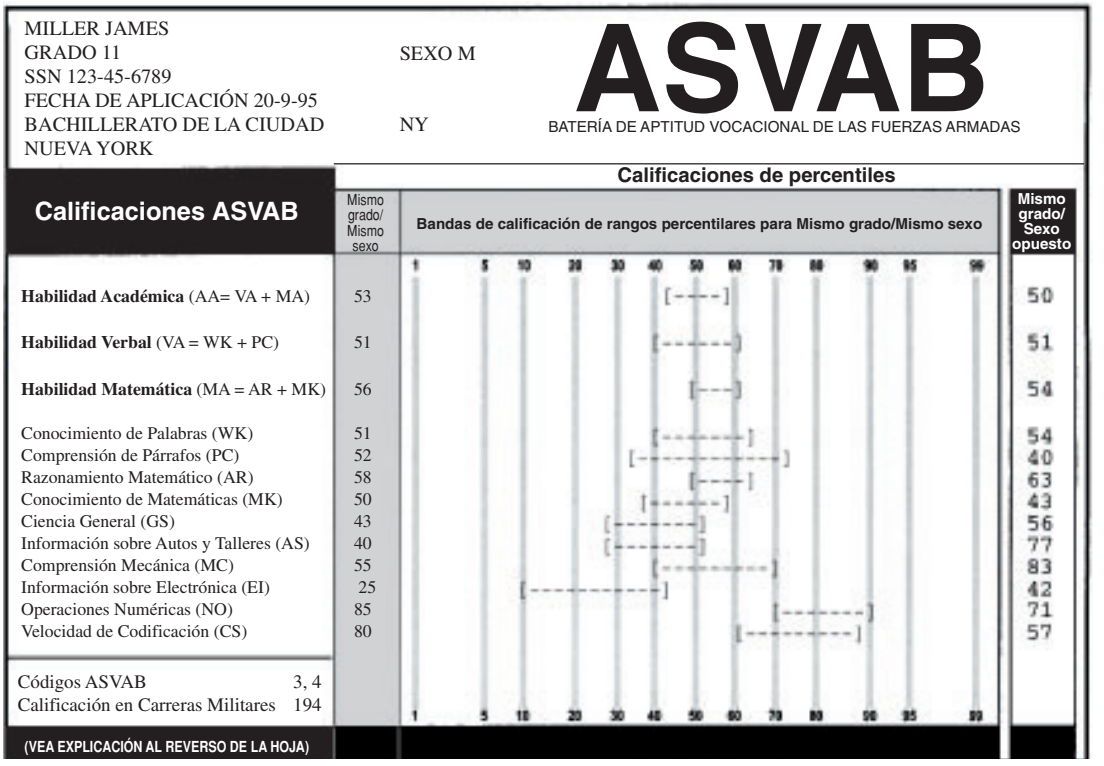


FIGURA 10.9 Perfil de Calificaciones en la Batería de Aptitud Vocacional de las Fuerzas Armadas.

(Reproducida con autorización del Departamento de Defensa de Estados Unidos a partir de *ASVAB 18/19 Counselor Manual.*)

luar sus posibilidades como candidato a las ocupaciones militares incluidas (vea Departamento de Defensa de Estados Unidos, septiembre de 1995).

Los coeficientes de confiabilidad por consistencia interna van desde los bajos .70 hasta los bajos .90 para las diez pruebas de ASVAB, y desde los bajos hasta los intermedios .90 para las tres calificaciones compuestas. Los coeficientes de confiabilidad de las formas opcionales están en su mayoría en los .70 y .80 para las diez pruebas, y en .90 para las calificaciones compuestas. En el *Manual Técnico para la ASVAB 18/19* (Departamento de Defensa de Estados Unidos, diciembre de 1999) se presenta gran cantidad de datos sobre la validez de la ASVAB para empleos tanto militares como civiles. Al igual que con la batería de las DAT, debido a que pueden aparecer las mismas pruebas en distintas combinaciones de ASVAB, estas combinaciones están positivamente correlacionadas. Así, la información proporcionada por diversas pruebas a menudo es redundante y refleja más la habilidad cognoscitiva que habilidades diferenciales específicas.

Además de una versión estándar en lápiz y papel, está disponible una versión de la ASVAB administrada por computadora (la CAT-ASVAB) que emplea una metodología de evaluación adaptativa. La CAT-ASVAB tiene las ventajas de un tiempo de administración menor, mayor seguridad, más precisión en la medición en los extremos de la habilidad, retroalimentación inmediata para los examinados sobre su desempeño, y tiempos de inicio flexibles (vea Segall y Moreno, 1999).

Evaluación diagnóstica y claves de trabajo

Desde su inicio en la década de 1920, la elaboración y el uso de pruebas de aptitudes múltiples se han basado en la suposición razonable de que distintos programas educativos y laborales requieren de distintas habilidades humanas. Idealmente, cuando se usan con propósitos de asesoría y colocación, dichos instrumentos de calificación múltiple cumplen la función diagnóstica de determinar los tipos de programas o empleos para los que la gente es más apta. En el futuro, sin duda, estas evaluaciones diagnósticas serán más adaptativas e individualizadas, y se aplicarán ya sea en una situación de interrelación entre personas o por computadora. Un posible escenario consiste en que el proceso de evaluación diagnóstica comience con una breve prueba de reto, se siga con pruebas para identificar las habilidades constitutivas en un área en que el examinado tenga problemas, luego de lo cual podrá iniciarse la construcción de un perfil de las aptitudes y deficiencias del examinando, y finalizar con la presentación de la instrucción o la capacitación necesarias como remedio. Las pruebas diagnósticas se propondrán, sobre todo, ayudar a los estudiantes y a los solicitantes de empleo a aprender y tener éxito, más que a simplemente proporcionar calificaciones para la toma de decisiones institucionales u organizacionales, y servirán como guía para la instrucción y capacitación en forma continua, más que sólo para comparar el desempeño de los examinados.

Durante muchos años las baterías de pruebas de aptitudes, tales como las DAT, la GATB y la ASVAB, se han administrado en contextos escolares y laborales para ayudar a los individuos a tomar decisiones educativas y vocacionales. Hay un procedimiento similar, pero más complejo y coordinado, el cual fue desarrollado por el American College Testing Program y pareciera de mayor utilidad. Este procedimiento, conocido como Sistema de Claves de Trabajo, consta de tres componentes o etapas: (1) un análisis de empleo o proceso de *elaboración de perfiles* con el cual determinar los niveles de habilidad requerida para un desempeño competente en trabajos específicos; (2) la evaluación de las habilidades de las personas en el lugar de trabajo, (3) la instrucción de apoyo para ayudar a los educadores a enseñar las habilidades requeridas. Durante la etapa 1, grupos intensivos concentrados y compuestos por trabajadores que realmente desempeñan un trabajo determinado son dirigidos por *personas dedicadas a elaborar perfiles* de trabajo para identificar la habilidad o habilidades requeridas en un empleo en particular. El resultado es

un perfil de trabajo que proporciona a empleadores, estudiantes, aspirantes a un empleo y a escuelas, un marco de referencia único para comprender cuáles son las habilidades que se necesitan para competir por determinado empleo. En la etapa 2, los alumnos o aspirantes que solicitan determinado empleo se someten a pruebas de dominio en cada una de las distintas áreas relevantes para el puesto: matemáticas aplicadas, tecnología aplicada, escuchar, ubicación de información, observación, lectura, trabajo en equipo y escritura. A continuación, las calificaciones individuales de estas evaluaciones se comparan con el perfil del trabajo en particular para revelar cualquier *laguna en las habilidades*. En la etapa 3, la información resultante obtenida en la etapa 2 se usa no sólo para proporcionar retroalimentación a quienes han realizado las pruebas, sino también para capacitarlos en las habilidades necesarias para el trabajo de que se trate. La capacitación necesaria para reducir las lagunas diagnosticadas en las habilidades consiste en instrucción basada en la computadora y de aula ajustada hacia metas específicas de Claves de Trabajo.

Además de funcionar como una base para la instrucción de reparación de lagunas, comparar las calificaciones de los estudiantes con los perfiles del trabajo también puede ayudar a los educadores a identificar los programas escolares que deben mejorarse. Asimismo, ejecutivos corporativos y administradores gubernamentales de desarrollo de la fuerza de trabajo pueden usar la información sobre las discrepancias entre calificaciones de evaluación de los empleados y los perfiles de trabajo para modificar sus programas de capacitación y contratación, y encauzar los fondos necesarios hacia las modificaciones requeridas (Doebele, 1999).

RESUMEN

Las pruebas de aptitudes o de habilidades especiales se centran en el futuro, es decir, en medir la habilidad para aprovechar de la capacitación adicional o de la experiencia en determinada área. Las pruebas de habilidades especiales también tienen amplitudes de banda más estrechas que las pruebas de inteligencia convencionales, en cuanto a que predicen logros más específicos. Aunque ciertas pruebas de habilidades especiales son del tipo de muestra de trabajo o de desempeño, se aplican con mayor frecuencia las pruebas de lápiz y papel.

Las pruebas que comprenden una batería de aptitudes múltiples se han estandarizado en la misma muestra de personas y, por ende, permiten comparar las diferencias existentes entre calificaciones individuales y calificaciones de varias personas. Las diferencias entre las calificaciones de una persona en una batería de aptitudes deberán interpretarse con precaución, y considerarse significativas sólo si son mayores que una o dos veces el error estándar de la medición de las diferencias entre calificaciones. Debido a que en una batería de aptitudes múltiples las pruebas, en general, son más cortas que las pruebas de capacidades específicas, las primeras suelen tener confiabilidades menores y, por lo tanto, errores estándar de medición mayores que las segundas.

La validez predictiva de las pruebas de habilidad es bastante baja en general, pero las calificaciones de tales pruebas pueden contribuir a predecir diversos criterios de desempeño cuando se usan en combinación con otras medidas de habilidad, así como con calificaciones previas sobre medidas de intereses, motivación y desempeño.

Existen algunas pruebas que tienen la finalidad de medir las agudezas visual y auditiva, la visión del color y otros aspectos relacionados con la sensación y la percepción. Además, se encuentran disponibles otros instrumentos de propósitos múltiples para la medición visual y de las habilidades perceptoras y motrices.

Las habilidades psicomotrices parecen ser altamente específicas y las calificaciones de pruebas de estas habilidads a menudo tienen menor confiabilidad que otras pruebas de habilidad. Las calificaciones de habilidades psicomotrices también son muy susceptibles a los efectos de la práctica. Como ejemplos de pruebas psicomotrices figuran la Prueba Minnesota de Índice de Manipulación para medir movimientos manuales gruesos, la Prueba Crawford de Destreza con Partes Pequeñas para medir movimientos manuales finos, y la Prueba Bennett de Destreza Mano-Herramienta para medir movimientos manuales tanto gruesos como finos.

Las pruebas de habilidades mecánicas y para el trabajo de oficina fueron de las primeras medidas estandarizadas de habilidades especiales en diseñarse. Sin embargo, ni la habilidad mecánica ni la habilidad para el trabajo de oficina constituyen una dimensión psicológica unitaria. Las pruebas de habilidad mecánica pueden incluir habilidades psicomotrices, además de la percepción y la comprensión mecánica.

Entre los ejemplos de pruebas de habilidad mecánica se encuentran la Prueba de Comprensión Mecánica Bennett y la Prueba de Conceptos Mecánicos. Las pruebas de habilidad para el trabajo de oficina pueden medir la velocidad y precisión perceptuales, así como la habilidad verbal y numérica. Las pruebas representativas de la capacidad para el trabajo de oficina incluyen la Prueba Minnesota de Trabajo de Oficina y la Batería de Capacidades para Trabajo de Oficina.

Se han diseñado varias pruebas con el propósito de medir las habilidades para la programación y operación de computadoras, incluyendo la Batería de Aptitud para Programador de Computadoras y la Batería de Aptitud para Operador de Computadoras.

Entre otras habilidades especiales para las que se han diseñado pruebas de habilidad se encuentra la de aptitud artística y musical. Sin embargo, la mayoría de las pruebas más antiguas de aptitud artística y musical ya no están disponibles comercialmente. Algunas de estas pruebas miden la apreciación artística (juicio y percepción), mientras que otras evalúan el desempeño artístico o el conocimiento sobre arte. Dos de las más populares han sido las Pruebas Meier de Percepción Estética y Juicio Estético, y la Prueba Graves de Juicio de Diseño.

Las Medidas Seashore de los Talentos Musicales, la prueba de habilidad musical más antigua publicada, ponen énfasis en la discriminación, el juicio y la memoria para las notas o combinaciones de notas. Otras pruebas de habilidad musical, por ejemplo el Perfil de Aptitud Musical, incluyen el juicio y la discriminación de música significativa. No obstante, el éxito ya sea en música o en arte depende de muchos factores adicionales al talento.

Las baterías de aptitudes múltiples se diseñan para medir las aptitudes y deficiencias en diversas áreas de habilidad. Las baterías de pruebas de aptitud, que por lo común no se aplican antes de los primeros años de bachillerato, son herramientas útiles en la asesoría, selección y colocación académicas y vocacionales. Ciertas baterías de habilidades múltiples, por ejemplo el Estudio de Habilidades Guilford-Zimmerman y la Batería de Pruebas de Aptitud General, se basan en los resultados del análisis factorial; no así otras como las Pruebas de Aptitud Diferencial y el Estudio de Aptitud del Empleado.

Las Pruebas de Aptitud Diferencial han sido una de las baterías más útiles para la asesoría académica, mientras que la Batería de Pruebas de Aptitud General se ha empleado más extensamente en consejería vocacional. La Batería de Aptitud Vocacional de las Fuerzas Armadas (ASVAB), la que más se ha aplicado de todas las baterías de aptitudes múltiples, se emplea para propósitos de colocación y selección ocupacional en el ejército estadounidense, y para asesoría de estudiantes de bachillerato que se interesan por carreras militares.

Las calificaciones de una batería de aptitudes, por sí solas, son inadecuadas para lograr una asesoría académica o vocacional efectiva. También deben tomarse en cuenta el desempeño pasado, los intereses y la motivación, las características de personalidad y todos los factores relacionados con la situación.

PREGUNTAS Y ACTIVIDADES

1. El apoyo empírico para distinguir entre aptitud y aprovechamiento se obtuvo en una investigación de corte transversal realizada por Burket (1973). Se encontró que las calificaciones de aprovechamiento *se elevaban* al aumentar el nivel de grado cuando las calificaciones de habilidad se mantenían constantes, pero las calificaciones de habilidad *disminuían* con el aumento del nivel de grado cuando las calificaciones de aprovechamiento se mantenían constantes. Estos hallazgos, combinados con los de otros investigadores (por ejemplo Carroll, 1973), pueden interpretarse en términos de la siguiente ecuación descriptiva: Aprovechamiento = Aptitud × Experiencia. Explique esto.

2. Identifique por lo menos dos pruebas en cada una de las siguientes categorías de habilidad: psicomotriz, espacial, mecánica, para el trabajo de oficina, artística, musical.

3. ¿Cuáles son las ventajas y desventajas de aplicar una batería de pruebas de aptitud en lugar de varias pruebas individuales de habilidades especiales?

4. ¿Cómo difieren los objetivos de la selección y el sondeo de personal de los de clasificación y colocación? ¿Qué tipos de pruebas son más adecuados para ayudar en el proceso de toma de decisiones para la selección y/o el sondeo? ¿En la clasificación y ubicación?

5. Escriba una reseña crítica sobre cualquiera de las pruebas individuales de habilidades especiales descritas en este capítulo; siga el lineamiento dado en el ejercicio 9 de la sección de Preguntas y Actividades del capítulo 6.

6. Haga una cita con un optometrista u oftalmólogo de su localidad y pídale que describa los procedimientos e instrumentos empleados para probar la visión de una persona. ¿Cuáles aspectos de la visión se miden de rutina y cuáles se miden sólo en circunstancias especiales? Prepare un informe sobre sus hallazgos.

7. Planee una visita a las oficinas administrativas de la escuela de su distrito y entreviste al psicólogo del plantel o al director de educación especial sobre las pruebas psicológicas aplicadas por ellos. Por ejemplo, ¿qué pruebas se usan para evaluar las habilidades o discapacidades especiales de los alumnos? ¿Con qué frecuencia se prueba o revisa la visión de los estudiantes para detectar problemas? Elabore un informe sobre sus hallazgos.

8. Juan obtiene una calificación *T* de 65 en la prueba de comprensión verbal y otra *T* de 75 en la prueba de razonamiento numérico de una batería de pruebas de aptitud múltiple. Si la confiabilidad de las dos pruebas es de .90 y .85, respectivamente, ¿puede el examinador estar 95% seguro de que Juan tiene más deficiencias en comprensión verbal que en habilidad numérica? Apoye su respuesta en cálculos adecuados.

9. ¿Cuáles son algunas de las diferencias de género en aptitudes o habilidades especiales que ha observado? ¿A qué factores atribuye estas diferencias?

CAPÍTULO ONCE

APLICACIONES Y PROBLEMAS EN LAS PRUEBAS DE HABILIDAD

La principal razón de que se apliquen pruebas de habilidades en escuelas, universidades y otras instituciones educativas es determinar la medida en que los estudiantes han acumulado conocimientos y habilidades específicos, ya sea dentro o fuera de ambientes académicos formales. El conocimiento debe incluir no sólo la simple repetición de hechos memorizados, sino también cierto grado de comprensión y capacidad para aplicar lo que se ha aprendido en varias situaciones y circunstancias. Del mismo modo, las habilidades aprendidas —cognoscitiva, psicomotriz y social— deben poder generalizarse o transferirse a otras áreas de la vida. La medición de estas habilidades involucra tanto a individuos (estudiantes, maestros, personal administrativo, etc.) como a grupos de personas (clases, escuelas, distritos escolares, muestras representativas de los residentes de estados y países) y los programas o procedimientos de intervención mediante los cuales se llevan a cabo cambios en conocimientos y habilidades.

Este último capítulo sobre la evaluación de habilidades empieza por considerar tres áreas en las cuales se ha concentrado la evaluación educativa en años recientes: la competencia de los estudiantes, la competencia de los maestros y programas de intervención. Un análisis de los esfuerzos concentrados en estas tres áreas debe proporcionar un panorama útil para conocer la manera en que se han administrado los instrumentos de evaluación psicológica con propósitos de evaluación y selección en escuelas, universidades y otras organizaciones. Desde luego, la administración de pruebas de habilidad tanto en ambientes educativos como en diversas instituciones no ha estado libre de críticas y controversias. Más que enterrar la cabeza en la arena y suponer arrogantemente que los críticos de las evaluaciones psicológicas y educativas sólo intentan llevar agua a su molino, es sabio científica, humanística y políticamente que los diseñadores y usuarios de pruebas presten atención, evalúen y escuchen bien estas críticas. Sólo así pueden esperar mejorar sus productos y servicios, y que resulten de mayor valor para la sociedad en su conjunto.

LA EVALUACIÓN EN EL CONTEXTO EDUCATIVO

Evaluación de la aptitud escolar

Apenas es de sorprender que los críticos de las escuelas públicas estadounidenses proclamen que las escuelas y los estudiantes de Estados Unidos están en problemas. Aunque más de tres cuartos de los adultos de ese país son graduados de bachillerato, los resultados de un cuarto de siglo de evaluaciones de la Evaluación Nacional del Progreso Educativo (NAEP) del conoci-

miento y las habilidades de los jóvenes revelan deficiencias persistentes en lectura, escritura, ciencia, matemáticas, historia, civismo y otras materias. Como se discutió en el capítulo 6, la medición periódica en los estudiantes de los niveles de habilidad en lectura, matemáticas, ciencia, escritura, historia, geografía y otras áreas académicas ha sido designada como La Tarjeta de Informe de la Nación. Los resultados de 20 años que se resumen en la tabla 11.1 indican que el aprovechamiento académico es bajo entre los alumnos blancos, negros y latinos por igual, en particular en los últimos dos grupos. Desde la década de 1970, los estudiantes negros y latinos de Estados Unidos han mejorado en lectura, matemáticas y ciencia, pero su desempeño sigue siendo considerablemente inferior al de los blancos.

Evaluación de la competencia académica de los estudiantes. La preocupación nacional sobre las bajas calificaciones en las pruebas de los graduados de bachillerato en Estados Unidos ha llevado a que en muchos estados se solicite que los estudiantes se sometan a una prueba de *alfabetismo funcional*, o de competencia mínima, antes de recibir un diploma de bachillerato. A pesar de los acuerdos y esfuerzos por volverla más aceptable, la evaluación de aptitud mínima o alfabetismo funcional ha sido objeto de continuos debates. Debido a que porcentajes considerablemente más altos de estudiantes negros que de blancos han fracasado en los exámenes estatales para estudiantes de bachillerato, a menudo se ha acusado a varias de estas pruebas de discriminar a las minorías (por ejemplo, *Debra P.* contra *Turlington*, 1984). También hay críticos que consideran el aprobar una prueba de octavo grado como un estándar inadecuado para la graduación de bachillerato, y que se corre el riesgo de convertir el desempeño mínimo en la norma. Dos peligros más de la evaluación de mínimo desempeño son que los maestros pueden terminar enseñando para la prueba y que quienes imponen la disposición seguirán siendo acosados por los indignados padres cuyos hijos fracasen en la prueba.

A pesar de estos problemas, en Estados Unidos el uso de pruebas para evaluar la competencia en habilidades básicas y la exigencia de calificaciones mínimas específicas para la graduación de bachillerato parece haber llegado para quedarse. En muchos estados, la representatividad me-

TABLA 11.1 Resultados más destacados en 20 años de NAEP

- Los estudiantes pueden leer a nivel superficial, obteniendo la esencia del material, pero no leen en forma analítica ni se desempeñan bien al realizar tareas de lectura con metas.
- Pequeñas proporciones de alumnos escriben lo bastante bien como para cumplir los propósitos de distintas tareas de escritura, pero la mayoría no se comunica en forma eficaz.
- Sólo reducidas proporciones de estudiantes desarrollan un conocimiento especializado necesario para abordar problemas basados en la ciencia, y el patrón de quedar rezagados se inicia en la escuela elemental.
- La adquisición de las cuatro operaciones aritméticas y del inicio de la resolución de problemas está lejos de ser universal entre los alumnos de la escuela elemental; para cuando se acercan a la graduación de bachillerato, la mitad no puede manejar material matemático de moderada dificultad, incluyendo cálculos con decimales, fracciones y porcentajes.
- Los alumnos tienen una comprensión básica de los acontecimientos históricos de Estados Unidos, pero parecen no entender su importancia y conexiones.
- De manera similar, los estudiantes demuestran una comprensión desigual de la Constitución del gobierno y la política estadounidenses; su conocimiento sobre el Acta de Derechos es limitado.

diante la evaluación del desempeño de los estudiantes es un acontecimiento anual que da como resultado una lista publicada en los periódicos locales de los porcentajes de calificaciones de la prueba por escuela y grado. Los esfuerzos (estadounidenses) por volver más útiles tales evaluaciones para la toma de decisiones escolares y la distribución de ingresos se manifiestan en las convocatorias para registrar calificaciones en las pruebas de NAEP por estado y localidad, más que por simples promedios de todo el país.

Evaluación de valor agregado. En Estados Unidos, el concepto de educación de valor agregado y el proceso asociado de la evaluación de valor agregado están relacionados con la representatividad y la evaluación de la competencia. En la *evaluación de valor agregado,* el aprovechamiento de los estudiantes en materias académicas y habilidades de la vida, tales como analizar una columna de periódico, una tabla matemática o un aviso publicitario televisivo se evalúan antes y después de cierto periodo de educación y estudio formal. La diferencia entre calificaciones de pruebas antes y después de cursos es una medida del valor agregado por la experiencia educativa. Por ejemplo, es posible pedir a los estudiantes de nuevo ingreso a la universidad que analicen publicidad, artículos y disertaciones de un periódico para demostrar su dominio de habilidades de la vida. Volver a aplicar la prueba al final del segundo año, cuando a los estudiantes aún les queda suficiente tiempo para compensar las deficiencias, revela cuánto han aprendido en el programa educativo general. La evaluación de valor agregado se exige por ley y la controlan consejos coordinadores de ciertos estados, e instituciones individuales en varios estados más han incorporado la evaluación de valor agregado en sus procedimientos académicos.

Maestros y evaluación

La evaluación en las escuelas se lleva a cabo por psicólogos, asesores y directores de educación especial, pero con mayor frecuencia por los propios maestros de aula. Desde su primer día en el salón de clases, los maestros se involucran en la evaluación formal e informal de los estudiantes. Tales evaluaciones implican no sólo observaciones, trabajo en clase, tareas en casa y pruebas elaboradas por maestros, sino también pruebas estandarizadas. Sin embargo, el amplio uso de las pruebas estandarizadas en las escuelas conduce, con frecuencia, a errores de administración, calificación e interpretación. Muchos de estos errores pueden atribuirse a falta de capacitación, de interés, o de ambos aspectos por parte de los usuarios de las pruebas. En consecuencia, es un asunto de cierta importancia que los maestros, los asesores y quienes tengan responsabilidades de evaluación en las escuelas estén adecuadamente capacitados e informados.

Capacitación de los maestros en evaluación. La mayoría de los prospectos de maestro tiene cierto contacto con evaluaciones psicológicas y educativas durante los cursos universitarios, pero en gran parte de los casos es bastante superficial. Muchos maestros no comprenden lo que miden las pruebas que están administrando; tampoco saben el significado de las calificaciones estándar que se inscriben en el registro permanente de los estudiantes. A menudo extraen conclusiones apresuradas con base en una única calificación de una prueba y no toman en cuenta la historia del desarrollo del niño, la competencia social o el ambiente familiar. Por lo tanto, es esencial que se preste más atención a este aspecto de la capacitación de los maestros. Por ejemplo, deben darse cuenta de que las calificaciones de las pruebas de inteligencia y habilidades especiales deberían interpretarse en términos de las probabilidades de que el examinando tendrá éxito en una vocación o programa de estudios en particular. Con demasiada frecuencia las calificaciones de pruebas se consideran medidas fijas de la situación mental, por una parte, o carentes de todo sentido, por la otra.

Evaluación de los maestros. El aumento del interés público sobre la calidad de la educación en Estados Unidos ha llevado a otra forma de involucramiento de los maestros en la evaluación. Casi todos los 50 estados han implantado algún tipo de sistema de evaluación de maestros. La prueba más usada para observar a los candidatos a maestros de nivel universitario y principiantes, y para certificar a los graduados en cuanto a conocimientos generales, habilidades profesionales y conocimiento de materias, es la Serie de Praxis descrita en el capítulo 6. De particular interés en el contexto actual es el desempeño de un candidato en Praxis III: Evaluaciones de Desempeño en el Aula, que consiste en un marco de capacitación y evaluación para las pruebas de aula.

La mayoría de los estados requiere una calificación aprobatoria en una prueba específica, como Praxis I, para que los alumnos ingresen a programas de capacitación para maestros, y casi todos los estados usan pruebas para la certificación de maestros. También se aplican pruebas con fines de recertificación y de asignación del pago justo. Además de la Serie Praxis y de otras pruebas, varios estados han instituido sistemas de observación formales para los maestros principiantes. En estos estados, los maestros inexpertos reciben colaboración en la enseñanza durante un periodo de prueba, al término del cual una recomendación propuesta a los funcionarios estatales determina si el candidato habrá de recibir la certificación formal.

Por desgracia, un gran porcentaje de posibles maestros no se desempeña bien en estas pruebas. Por ejemplo, en muchos estados un tercio o más de los individuos que se someten a la Praxis I no alcanzan la calificación eliminatoria establecida para maestros principiantes. Asimismo, como grupo, los estudiantes que afirman se graduarán en pedagogía obtienen calificaciones menores que el promedio en pruebas de admisión a la universidad como la SAT y la ACT. Las bajas calificaciones en candidatos y maestros practicantes se atribuyen, al menos en parte, al hecho de que la enseñanza se ha vuelto menos atractiva para mujeres competentes y minorías en comparación con otras vocaciones más lucrativas y prestigiosas.

Los resultados de encuestas de opinión nacionales indican que la mayoría del público general está en favor de usar pruebas de competencia para la certificación y autorización de maestros (Gallup, 1991). Además, las dos mayores organizaciones de maestros del país, la Asociación de Educación Nacional y la Federación Americana de Maestros, apoyan la evaluación de maestros principiantes para garantizar que cumplan con un estándar de aptitud razonable. Los partidarios de una prueba nacional para candidatos a maestros sostienen que sería un indicador de la calidad de los maestros y volvería profesional al gremio. Al mejorar la calidad de los maestros, dicha prueba también apoyaría los aumentos salariales para ellos, así como mejoras generales en la calidad de las escuelas.

Las pruebas de aptitud para maestros no han carecido de retos y en varios estados se han librado batallas legales concernientes a dichas pruebas. Un problema constante se refiere al estándar aprobatorio: si se establece en un nivel razonablemente alto, entonces una gran cantidad de candidatos minoritarios probablemente fracasarán; si el nivel determinado es muy bajo, individuos con baja habilidad ingresarán a la profesión de maestros. Asimismo, ciertos educadores profesionales han expresado su desacuerdo con el carácter de los exámenes. Algunas autoridades consideran que una mezcla de pruebas con el uso de tecnología de computación, observaciones directas del desempeño en el aula, un portafolio con documentación sobre desempeño pedagógico y otros requisitos, así como pruebas estandarizadas de lápiz y papel, deberían emplearse para evaluar tanto a los futuros maestros para su contratación, como a los maestros con experiencia para recertificación, promoción, cargos y pago justo.

Evaluación de programas

Además de evaluar la aptitud de alumnos y maestros, suelen usarse pruebas, escalas de evaluación y cuestionarios para medir la eficacia de los programas educativos y otras intervenciones.

La evaluación psicológica y educativa desempeña un papel importante al juzgar la instrucción y determinar la efectividad de los tratamientos psicológicos y otros procedimientos diseñados para modificar comportamientos, cognición y actitudes. Tales programas no deberían diseñarse unilateralmente por especialistas en psicometría e investigación, sino en colaboración con educadores, personal de servicios humanos, personal de salud, funcionarios públicos y otros profesionales del área de intervención. Sin embargo, las contribuciones de especialistas en medición resultan ser las más importantes para recomendar y/o diseñar instrumentos con qué evaluar los resultados de programas.

Las dificultades para medir el cambio y otros problemas técnicos de evaluar la efectividad de las intervenciones sobre el comportamiento han conducido a la creación de un nuevo tipo de especialidad: la evaluación de programas. Según la definen Posavec y Carey (1997), la *evaluación de programas* es

> una colección de métodos, habilidades y sensibilidades para determinar si un servicio humano es necesario y si es factible usarlo, si es lo bastante intenso como para cumplir con la necesidad identificada no resuelta, si el servicio se ofrece como se planeó y si efectivamente ayuda a las personas que lo necesitan sin efectos secundarios indeseables (p. 1).

El objetivo de la evaluación de programas es emitir juicios relativos a la utilidad o el valor de programas educativos, psicosociales y otros programas de intervención social. Se han propuesto diversas guías o modelos de evaluación de programas, incluyendo el modelo CIPP (contexto, entrada, proceso, producto), la evaluación de discrepancia y la evaluación adversaria. Se han escrito muchos libros y artículos sobre el tema de la evaluación de programas, pero aquí sólo se presentará una breve descripción de los métodos.

Posavec y Carey (1997) describen la filosofía y los objetivos de la evaluación de programas en términos de necesidades, proceso, resultado y eficiencia. Primero se evalúan las *necesidades* de las personas para quienes una organización podría proporcionar un servicio. A continuación, se establece un programa diseñado para cubrir dichas necesidades, y se vigila el *proceso* mediante el cual se aplicará para determinar si se requieren ajustes. Después que el programa ha estado funcionando por un tiempo, se revisan los *resultados* (el grado en que el programa ha sido efectivo para cumplir sus metas). Además de estimar la eficacia del programa, se evalúa su *eficiencia*, es decir, el costo monetario en relación con los resultados. En este punto se toma la decisión de continuar, descontinuar o modificar el programa y/o sus objetivos de alguna manera.

Rossi y Freeman (1993) propusieron un modelo inclusivo similar para la evaluación de programas. Este modelo caracteriza el proceso general de evaluación de programas en términos de cuatro etapas sucesivas: planeación, monitoreo, evaluación de efectos y evaluación de la eficiencia económica. Durante la primera etapa, o de *planeación del programa*, se identifican el alcance del problema (por ejemplo, tráfico y uso de drogas en las escuelas), los objetivos y la población meta del programa. Después de haber especificado los objetivos y la población meta, se toma una decisión en cuanto a si el programa puede aplicarse de manera apropiada. Una vez que se ha decidido continuar, comienza la etapa de *monitoreo del programa*. Entonces, la aplicación o el funcionamiento del programa se vigilan en forma continua en cuanto a si proporciona los recursos y servicios designados a la población meta.

En la tercera etapa, o de *evaluación de efectos*, los resultados reales se evalúan para comprobar que se hayan satisfecho los objetivos del programa. Se emplean diversos procedimientos estadísticos y no estadísticos para determinar si los resultados son significativos y si se encuentran en la dirección pronosticada. Por ejemplo, los criterios de efectividad de un programa de tratamiento

psicológico pueden incluir evaluaciones de la importancia de los cambios, la proporción de individuos que mejoran, el alcance de los cambios y la durabilidad de la mejoría (Kazdin, 1998).

En la etapa de evaluación de efectos se miden también otros resultados no planeados o inesperados, pero incluso cuando sean estadísticamente significativos, pueden no tener la suficiente importancia práctica como para garantizar la aplicación del programa. Por consiguiente, el propósito de la cuarta etapa, *evaluación de la eficiencia económica*, es determinar si los resultados de un programa valen los costos generados. Al evaluar la eficiencia de un programa de tratamiento, por ejemplo, deben considerarse factores tales como la duración del tratamiento, su difusión y sus costos monetarios (Kazdin, 1998). La evaluación de la eficiencia económica es un asunto de análisis de costo-beneficio, en el cual los gastos del programa se comparan con sus beneficios potenciales para el individuo y la sociedad. Incluso si el programa funciona, es posible que los recursos monetarios y de otro tipo necesarios para ponerlo en práctica se usen de manera más efectiva en otros fines. Cuando los resultados de un análisis de costo-beneficio favorecen el programa, es una señal para seguir adelante y ponerlo en funcionamiento. Pero antes de tomar la decisión final de extender el programa por más tiempo y a otros contextos, es sensato definir su aceptabilidad para quienes, directa o indirectamente, resultan afectados por él. Un programa educativo y social puede tener ramificaciones tanto políticas como personales y sociales con respecto a su aceptabilidad para una porción más amplia de la sociedad, y no sólo para quienes estuvo explícitamente diseñado. Incluso después de iniciado el programa, su eficacia debería evaluarse y revisarse periódicamente.

Aunque varios modelos de la evaluación de programas difieren en los detalles, todos intentan determinar los objetivos, recursos, procedimientos y administración del programa con el fin de juzgar su mérito. Como indicio del nivel de interés en estos esfuerzos, y del apoyo público hacia ellos, están los centros de investigación y desarrollo en evaluación educativa y otros tipos de evaluaciones de programas en destacadas universidades estadounidenses. Los hallazgos de los estudios realizados en estos centros contribuyen a proporcionar una base más racional para responder las preguntas sobre los procesos y resultados de diversos tipos de programas sociales.

CRÍTICAS Y PROBLEMAS EN LAS PRUEBAS DE HABILIDAD

Como lo muestran la cantidad y diversidad de instrumentos descritos en los cinco capítulos precedentes, la evaluación de capacidades cognoscitivas, perceptuales y psicomotrices se expandió rápidamente durante el siglo XX. La extensa aplicación de pruebas colectivas de aprovechamiento, inteligencia y habilidades especiales en educación, los negocios y el gobierno ha contribuido al desarrollo de la evaluación psicológica de empleados. No obstante, la mano de obra organizada, sosteniendo que la selección y la promoción laboral deberían basarse en la experiencia y la antigüedad antes que en calificaciones de pruebas, en general no ha apoyado las evaluaciones psicológicas. También se ha declarado una abierta oposición a las pruebas estandarizadas en contextos educativos, en particular al uso de exámenes de admisión a la universidad y a las pruebas de inteligencia aplicadas en las escuelas.

Encuesta Phi Delta Kappa

Los estudios anuales sobre las actitudes de los estadounidenses frente a la evaluación en las escuelas han revelado una controversia cada vez mayor con respecto al uso de pruebas estandarizadas en decisiones arriesgadas que conciernen a los alumnos. Por ejemplo, en la XXXIII

Encuesta Anual Phi Delta Kappa/Gallup (Rose y Gallup, 2001), 31% de los encuestados respondieron que había demasiado énfasis en la evaluación del aprovechamiento en las escuelas. Porcentajes todavía más elevados se opusieron al uso de una sola prueba estandarizada para determinar si un estudiante debía ser promovido de un grado a otro (45%) y para decidir si un alumno debería recibir un diploma de bachillerato (42%). Se opusieron a estos usos de pruebas estandarizadas porcentajes más altos de encuestados de 18 a 29 años de edad que mayores de 65 años, porcentajes más elevados de negros que de blancos, y mayores porcentajes de demócratas que de republicanos. Aproximadamente dos tercios de todos los encuestados consideraron que las pruebas deberían usarse principalmente para establecer el tipo de instrucción requerida más que para determinar cuánto habían aprendido los alumnos, y que el trabajo en el aula y en casa eran mejores parámetros del aprovechamiento académico que las calificaciones de pruebas.

Carácter y consecuencias de las críticas

La mayoría de las críticas a la evaluación psicológica y educativa durante las últimas décadas se ha ocupado ya sea del contenido y los usos de las pruebas o bien de las consecuencias sociales de confiar en calificaciones de pruebas para tomar decisiones sobre la gente. Se ha atacado a las evaluaciones en general, por una parte, por invadir el derecho del individuo a la intimidad y, por otra parte, por su secreto o confidencialidad. Las pruebas de habilidad, en particular, se han visto acusadas de tener limitaciones y sesgos en lo que intentan medir.

Con respecto a sus usos, se ha argumentado que, más que propiciar la igualdad de oportunidades, las pruebas han provocado la conservación del estado de cosas y la legitimación de prácticas antidemocráticas por parte de instituciones educativas, organizaciones empresariales y el propio gobierno. De manera más específica, se ha sostenido que las pruebas a menudo resultan inútiles para predecir el comportamiento, son injustas con los grupos minoritarios, suelen malinterpretarse y sus resultados se utilizan de modo inadecuado, promueven una clasificación de las personas estrecha y rígida de acuerdo con características supuestamente estáticas.

Las críticas a la evaluación psicológica y educativa con frecuencia sólo han provocado ruido y poco esclarecimiento, pero algunas de las preocupaciones han propiciado que se reconsideren las prácticas de evaluación. Ciertas críticas han originado cambios de carácter técnico, mientras que otras han impulsado un nuevo examen de la ética de las evaluaciones, así como el esbozo de propuestas de un código de ética que sería válido para editores, distribuidores y usuarios de las pruebas.

Los problemas legales y éticos relacionados con la aplicación de pruebas psicológicas y el uso de resultados de pruebas se discutieron brevemente en el capítulo 1. Como se señaló ahí, de acuerdo con el Acta Familiar de los Derechos Educativos y de Privacía (1974), las calificaciones de pruebas y sus interpretaciones que conservan las instituciones educativas pueden estar disponibles para otras personas sólo con el *consentimiento informado* del alumno o de un adulto legalmente responsable del mismo. Pero, incluso cuando se ha otorgado consentimiento informado, los datos de pruebas pueden ser *privilegiados* en cuanto a que sólo ciertas personas (padres, abogado personal, médico, psicólogo y otros especialistas) tienen derecho de acceso a ellos.

El concepto de comunicación privilegiada también se aplica a la información de pruebas y de otro tipo. Sin embargo, la información privilegiada es un asunto de todo o nada: un psicólogo que esté autorizado por el cliente para revelar información específica relativa a un caso, debe revelar *toda* la información disponible que sea relevante para el caso cuando así se lo ordene una corte. Asimismo, siempre que un psicólogo piense que un cliente representa un peligro claro y real para sí mismo o para otros, puede entregar la información privilegiada a personas res-

ponsables sin el consentimiento del cliente. De hecho, debido a que el bienestar de la sociedad en su conjunto se antepone al derecho de un individuo a la intimidad y a la comunicación privilegiada, los psicólogos pueden estar legalmente obligados a revelar la información (*Tarasoff* versus *Regents of University of California*, 1983).

Se ha discutido ampliamente si la aplicación de pruebas psicológicas representa una invasión grave a la intimidad. Puede argumentarse que si las respuestas a las preguntas de la prueba tienen suficiente valor social, entonces el individuo tendrá que soportar cierta invasión a su intimidad. Por importante que pueda ser el respeto a los derechos individuales con respecto a la confidencialidad de las calificaciones de pruebas y la invasión a la intimidad, estos derechos deben ponerse en equilibrio frente a la necesidad de la sociedad de contar con información de evaluación de alta calidad.

De manera ideal, los resultados de las evaluaciones psicológicas se manejan conscientemente y tomando en cuenta las limitaciones del instrumento y las necesidades y derechos de los examinados. Desafortunadamente, los estándares éticos de los examinadores no son siempre tan elevados como deberían. La conciencia de este problema condujo a la Asociación Psicológica Estadounidense y a otras organizaciones profesionales a adoptar códigos de ética relativos a las evaluaciones, y a imponer sanciones en contra de la violación de estos códigos (American Psychological Association, 1981, 1992; American Educational Research Association *et al.*, 1999). Esto representa un paso adelante en la evaluación psicológica y la práctica de la psicología en general.

Exámenes de admisión a la universidad

Los programas de evaluación a gran escala, en donde se aplican pruebas a miles de estudiantes cada año, han sido objeto especial de crítica durante las últimas décadas. Por ejemplo, se ha sostenido que se dedica demasiado tiempo escolar a administrar pruebas que sólo miden algunas variables pertinentes para el aprovechamiento académico y otros logros. De todos los programas de evaluación a gran escala, los de mayor influencia y que se atacan más a menudo son los que incluyen exámenes de admisión universitarios. La Prueba de Evaluación Académica (SAT), las Pruebas Universitarias Estadounidenses (ACT), y varios otros instrumentos caen en esta categoría, pero la SAT ha sido objeto de las críticas más implacables.

Es probable que la mayoría de los funcionarios de la admisión universitaria asignen más peso a los grados de bachillerato y las calificaciones de la SAT que a indicadores del desempeño tales como entrevistas orales, cartas de recomendación, actividades extracurriculares y revisión de trabajos. Esto es comprensible cuando consideramos la baja objetividad y escasa confiabilidad de muchas de estas medidas de "cualidades personales" y desempeño. Por ejemplo, debido a la falta de confiabilidad o a la preocupación al respecto y un intenso interés por parte de quien escribe la carta de que se acepte al candidato, las cartas de recomendación casi siempre son laudatorias. Por esta razón, se ha afirmado que "una llamada telefónica vale una docena de cartas de recomendación". El mismo error de indulgencia, además de la variabilidad en cuanto a los estándares de los grados de una escuela a otra, afectan la precisión de los grados de bachillerato para pronosticar el desempeño en la universidad. Las entrevistas personales siguen teniendo cierto valor para las admisiones, pero también están limitadas por los prejuicios del entrevistador y la habilidad de los solicitantes para presentarse a sí mismos en forma efectiva.

A pesar de que pocas universidades requieren se anexen calificaciones de la SAT a las solicitudes, la gran mayoría de estas instituciones ha conservado ya sea la SAT o las ACT con propósitos de admisión y colocación. Las calificaciones de estas pruebas también pueden funcionar como un sistema de advertencia primario y como guías de diagnóstico para el trabajo de actualización. La SAT es una de las pruebas más cuidadosamente diseñadas de todas las que se encuentran disponibles, y tiene una elevada confiabilidad y considerable validez para predecir

grados universitarios. Sin embargo, estos rasgos no la han protegido de la ola de críticas a que ha estado sometida desde la década de 1950. La SAT, así como otras medidas psicométricas de la esperanza académica y el progreso, a menudo han funcionado como chivo expiatorio para ocultar las desventajas del sistema educativo en su conjunto.

Pruebas de opción múltiple

Durante la década de 1960, los críticos de los exámenes de admisión a la universidad y de otras pruebas educativas administradas en todo Estados Unidos (por ejemplo, Black, 1962; Hoffman, 1962) estuvieron especialmente activos. De estos críticos, el más estridente y de mayor influencia fue Banesh Hoffman, quien argumentó que las pruebas de opción múltiple (1) propician lectores astutos, ingeniosos y rápidos; (2) penalizan a las personas sutiles, creativas y más profundas; (3) se interesan sólo por la respuesta y no por la calidad de la reflexión en que se basa o la habilidad con que se expresa, y (4) en general tienen un mal efecto en la educación y el reconocimiento del mérito. Sin embargo, estas acusaciones sólo se basaban en ejemplos hipotéticos y argumentos cargados emocionalmente más que en pruebas sólidas.

Las críticas de Hoffman y de los demás autores no quedaron sin respuesta. Tras examinar los supuestos básicos de varios críticos de la evaluación educativa, Dunnette (1963) concluyó que la mayoría de dichos supuestos eran erróneos y falaces debido a una falta de información o al rechazo a reconocer que las pruebas son las medias disponibles más precisas para identificar el mérito. Otras autoridades (por ejemplo, Chauncey y Dobbin, 1963) admitieron que las pruebas tienen limitaciones pero que, cuando se usan en forma apropiada, pueden ayudar a mejorar la enseñanza.

Los ataques a las pruebas estandarizadas no desaparecieron con la década de 1960, ni tampoco se limitaron a no psicólogos. Por ejemplo, el prominente investigador en psicología David McCleland (1973) argumentaba en favor de que se descontinuara el uso de todas las pruebas de opción múltiple. Sostenía que era preferible desarrollar otras mediciones, tales como las que evalúan la habilidad de aprender rápidamente, más que continuar usando medidas de lo que una persona ya sabe como forma de demostrar sus habilidades.

Una crítica de las pruebas de opción múltiple que es difícil de probar o refutar, pero que tiene amplias implicaciones educativas y sociales, sostiene que tales pruebas no sólo son medidas deficientes de la habilidad y el aprovechamiento, sino que también fomentan una enseñanza inferior y hábitos de estudio inadecuados. Ya sea que esta crítica esté o no justificada, se recomienda a los maestros cuidarse de no confiar demasiado en las pruebas objetivas, y que no pasen por alto los exámenes tradicionales de ensayo donde se exige a los alumnos que expliquen y apoyen sus respuestas (vea Courts y McInerney, 1993; Gifford y O'Connor, 1992). El uso efectivo de reactivos de ensayo requiere que quienes califican evalúen no sólo el contenido de las respuestas, sino también el estilo o la habilidad con que se expresan. Escribir la respuesta a una pregunta no mejora la habilidad para expresarse por medio de la escritura a menos que se proporcione una retroalimentación constructiva sobre la forma y el contenido de la respuesta.

La crítica de que las pruebas de opción múltiple brindan tan sólo un vistazo del conocimiento del estudiante a un nivel superficial y no logran revelar lo que puede hacer el alumno con ese conocimiento ha impulsado un movimiento hacia la *evaluación basada en el desempeño*, o *evaluación auténtica*, en las escuelas públicas. Consistentes en preguntas abiertas y resolución de problemas prácticos en ciencia matemática y en algunas otras materias, las pruebas basadas en el desempeño someten a esfuerzo al razonamiento, el análisis y la escritura. En dichas pruebas, los estudiante obtienen créditos no sólo por dar la respuesta correcta sino por demostrar cómo llegaron a ella. También puede solicitarse a los estudiantes que trabajen en grupos pequeños, realicen experimentos y compartan sus interpretaciones de los resultados, o que produzcan algo mediante el esfuerzo colec-

tivo. También puede evaluarse un conjunto de habilidades o productos de los estudiantes durante un periodo determinado, un proceso conocido como *evaluación de portafolio*. A pesar del entusiasmo de contar con nuevas pruebas, quedan por resolverse los problemas de validez, justicia, relación costo-beneficio y confiabilidad de las calificaciones con respecto a las evaluaciones basadas en el desempeño (Baker, O'Neil y Linn, 1993; Educational Testing Service, 1992).

Nuevos ataques contra el Servicio de Evaluación Educativa

Durante la década de 1980, la campaña más publicitada contra las pruebas estandarizadas y los exámenes de admisión a la universidad en particular, fue dirigida por el defensor de consumidores Ralph Nader y sus "soldados". En discursos e informes escritos, Nader criticaba las pruebas SAT, GRE y LSAT y otras pruebas de habilidad estandarizadas por no medir la imaginación, el idealismo, la determinación y otros atributos humanos que consideraba importantes para el progreso de la civilización. Nader sostenía que el uso de estas pruebas había provocado la restricción de las opciones de carrera de los estudiantes y el desperdicio de una gran cantidad de talento profesional.

Allan Nairn (Nairn y Asociados, 1980), un socio de Nader, sostenía que las calificaciones de la SAT y otras pruebas del Servicio de Evaluación Educativa (ETS) clasifican a las personas por clase social más que por habilidad, un hecho del que Nairn acusó a ETS de tratar de suprimir. El resultado, alegaba Nairn, es la negación de oportunidades educativas a estudiantes de nivel socioeconómico inferior y, por lo tanto, la conservación del estado de cosas en la educación superior. Nairn también concluyó que la prueba SAT resulta deficiente para predecir los grados universitarios y que debería abandonarse en favor de varias medidas diagnósticas de habilidad y competencia. Solicitó que se revelaran totalmente las preguntas y respuestas de la SAT y se admitiera que la prueba no mide ningún concepto tan general como la "habilidad académica".

El ETS respondió extensamente al ataque de Nader y Nairn (Educational Testing Service, 1980a, 1980b) concluyendo que las pruebas no niegan oportunidades a niños de familias pobres o de clase trabajadora, y que la SAT en particular no es deficiente para predecir el desempeño académico. Los funcionarios del ETS admitieron que ninguna prueba es capaz de pronosticar en forma perfecta el éxito ni académico ni en la vida, y que tampoco es una medida del valor o mérito de una persona. La SAT y otras pruebas de habilidad académica nunca tuvieron la intención de medir la habilidad innata, buscaron más bien evaluar las habilidades aprendidas en una amplia gama de actividades de tipo escolar.

El ataque de Nader y Nairn contra el ETS fue ampliado por el Centro Nacional para la Evaluación Justa y Abierta (FairTest). FairTest mantuvo el argumento de que los reactivos de la prueba SAT a menudo están sesgados y son injustos para grupos minoritarios y mujeres y que, consecuentemente, las pruebas privan a estos grupos de oportunidades educacionales equitativas. Otra preocupación que expresaba FairTest era que no resultaba ético pedir a los alumnos resolver secciones experimentales de la SAT, la GRE y otras pruebas del ETS consistentes en reactivos que no se califican pero se usan con propósitos de ensayo. FairTest demandó al ETS obtener el consentimiento de los examinados antes de hacerlos resolver secciones experimentales de la SAT. El Acta de Derechos de FairTest también destaca que los examinados tienen derecho a recibir información segura sobre la resolución de las pruebas y consejos acerca de estrategias; pruebas cronometradas con precisión y aplicadas en condiciones tranquilas; confidencialidad de las calificaciones y otros datos personales; el proceso debido para cualquier reto de la prueba, y acceso a los datos sobre la precisión de ésta (Weiss, Beckwith y Schaeffer, 1989).

Los estudiantes y sus padres tienen el derecho legal a la información concerniente al desempeño del estudiante en pruebas educativas o psicológicas, pero esto no necesariamente significa que las calificaciones reales deban revelarse. Más bien, los resultados de las pruebas deben

comunicarse de tal modo que no se malinterpreten o se les dé un mal uso y que ayuden antes que poner obstáculos a los estudiantes. Esta advertencia se aplica sobre todo a las pruebas aplicadas a niños con propósitos diagnósticos en contextos clínicos o educativos. Por otra parte, las calificaciones de exámenes de admisión a la universidad se comunican de rutina a los examinados así como a las instituciones que los estudiantes indican previamente. Además, la ley de Nueva York sobre veracidad en las evaluaciones, puesta en práctica en 1979, exige que a los estudiantes que se someten a la SAT o a otras pruebas de admisión a la universidad se les entreguen copias de las preguntas reales y las respuestas correctas, así como copias de sus propias hojas de respuestas, en un periodo razonable posterior a la prueba. Dos disposiciones más de la ley del estado de Nueva York son que (1) en el momento de la aplicación se comunique a quienes se sometan a la prueba cómo se calcularán sus calificaciones, cuál es la obligación contractual que el examinador tiene hacia ellos, y cómo las calificaciones de la prueba pueden ser afectadas por la asesoría y diversos factores demográficos, y (2) el concesionario de la prueba debe archivar la información y los estudios sobre su validez ante la comisión de educación estatal. La ley también exige que se publiquen ediciones completas de las pruebas para que los estudiantes puedan practicar con ellas.

Algunos críticos de la evaluación educativa desean ampliar las disposiciones de la ley de Nueva York, acerca de la revelación total, hacia otros estados, e incluir otros exámenes a fin de promover el uso de pruebas nuevas para disminuir el sesgo cultural y que la industria de la evaluación resulte más confiable para los consumidores. Aunque más de 24 legislaturas estatales, así como el gobierno federal, han considerado leyes similares a la del estado de Nueva York, el único otro estado que aplica un estatuto especial de regulación de exámenes de admisión a la universidad es California. Esta ley, conocida como Acta Dunlop, requiere sólo que se proporcionen muestras representativas de las pruebas al Departamento de Educación del Estado de California. Las legislaturas estatales de Nueva York y California han considerado una legislación adicional para volver más estrictas las reglamentaciones relativas a la evaluación, pero dichos esfuerzos sólo han tenido éxito en Nueva York.

El estatuto del estado de Nueva York y otras *legislaciones sobre la veracidad en las evaluaciones* en trámite no afectan únicamente a las pruebas SAT, ACT y a otras pruebas de admisión a la universidad, sino también a pruebas de admisión a escuelas de posgrado y profesionales. Aunque el Consejo de Admisión de la Escuela de Leyes y el Consejo de Admisión de Administración de Graduados aprobaron la revelación de los resultados de sus pruebas (LSAT y GMAT), la Asociación Estadounidense de Escuelas Médicas y la Asociación Dental Estadounidense expresaron una enérgica oposición a la legislación sobre la veracidad en la evaluación. La primera organización, argumentando que la ley de Nueva York viola los derechos de autor sobre la MCAT, obtuvo un interdicto en 1979 contra la aplicación de la ley. En 1990, una corte federal estableció que el estatuto del estado de Nueva York, que exige la publicación de los materiales de la Prueba de Admisión a la Escuela Médica, viola la ley federal de derechos de autor. A pesar de esta reglamentación, la revelación de materiales de prueba sigue siendo una práctica común en las organizaciones de evaluación. Los procedimientos actuales diseñados para garantizar evaluaciones justas y abiertas son una parte aceptada de la elaboración, administración y calificación de pruebas en el Servicio de Evaluación Educativa, el Programa de Evaluación Universitaria Estadounidense y otras organizaciones que diseñan y distribuyen pruebas.

Las preocupaciones en torno a la legislación sobre la veracidad en la evaluación han propiciado mejoras en la revisión en cuanto a que las preguntas de las pruebas no contengan sesgos culturales o socioeconómicos. La cuidadosa revisión interna llevada a cabo por el personal profesional del ETS ha eliminado los sesgos (de grupo étnico, género, etc.) de casi todos los miles de reactivos que incluyen las pruebas del ETS cada año. Además, el Consejo de Exámenes de Admisión a la Universidad ha adoptado la política de permitir que los estudiantes verifiquen sus

calificaciones de la SAT y que se presenten públicamente los reactivos de esta prueba un año después de haberse aplicado. Los examinados pueden también confrontar los reactivos de la SAT y de otras pruebas del ETS y la forma en que se aplican estos exámenes.

Efectos de la asesoría en las calificaciones de prueba

Es comprensible que los candidatos a ingresar a universidades de licenciatura y posgrado y a otras escuelas profesionales estén interesados en mejorar sus calificaciones en los exámenes de admisión. Como consecuencia de la creciente importancia de la evaluación nacional a gran escala, se han publicado folletos de asesoría de pruebas y establecido escuelas que aseguran poder aumentar la calificación de una persona en una prueba en particular o de pruebas estandarizadas en general. Tres de estas organizaciones de asesoría de pruebas son College PowerPrep, Kaplan Inc., y The Princeton Review.

El que la asesoría tenga o no un efecto significativo en las calificaciones de la SAT y en otros exámenes de admisión ha sido un tema discutido durante muchos años. Es un problema importante, pues si se demostrara que la asesoría puede mejorar las calificaciones de las pruebas, entonces los jóvenes que no pudieran pagar dicha asesoría carecerían de las mismas oportunidades que sus compañeros más pudientes.

Los resultados de los primeros estudios sobre asesoría indicaron que sus efectos varían ampliamente, dependiendo de la semejanza del material estudiado con el de la prueba, del nivel de motivación y educación del examinando y de otros factores. Hace algunos años el Consejo de Exámenes de Admisión a la Universidad (1971) presentó pruebas relativas a los efectos de la asesoría para la SAT. Los resultados indicaron que el estudio intensivo de corto plazo sobre reactivos similares a los de la SAT no produjo aumentos significativos de las calificaciones, especialmente en la sección verbal de la prueba. Sin embargo, esta conclusión fue puesta en entredicho por varias personas, en particular por Stanley H. Kaplan, director de la mayor organización de asesoría de pruebas en el mundo. En 1979, la Comisión Federal de Comercio (FTC) presentó el informe de un estudio sobre los efectos de un programa de asesoría de diez semanas llevado a cabo en tres de los centros educativos Kaplan. Admitiendo que el estudio adolecía de ciertas fallas metodológicas, la FTC concluyó no obstante que el desempeño en las secciones verbal y matemática de la SAT mejoró gracias a los cursos de asesoría.

El estudio de la FTC y una revisión de los resultados efectuada por Slack y Porter (1980) se evaluaron posteriormente por el Servicio de Evaluación Educativa. Al analizar de nuevo los datos de la investigación de la FTC, el ETS obtuvo resultados similares: efectos inconsistentes e insignificantes de la asesoría para estudiantes en dos de las escuelas Kaplan, e incrementos de 20 a 35 puntos para calificaciones en las secciones verbal y matemática en una tercera escuela. A pesar de reconocer que puede haber aumentos considerables en las calificaciones cuando los programas de asesoría incluyen muchas horas de trabajo en los cursos y tareas, el ETS afirmó que por lo menos parte de los aumentos descubiertos en la tercera escuela podrían atribuirse a diferencias en la motivación y a otras características personales.

Los resúmenes de los estudios realizados durante las últimas dos décadas sobre los efectos de la asesoría en las calificaciones de la SAT revelan que el estudio intensivo de reactivos similares a los de la prueba puede producir aumentos de 15 a 25 puntos en las secciones tanto verbal como matemática. Sin embargo, estos aumentos no son mayores que los observados en estudiantes que repiten la prueba después de otro año de bachillerato (Donlon, 1984). Las mejoras ocurren sobre todo en reactivos con formatos complejos o confusos y con individuos de contextos educativos deficientes (Powers, 1986). Acertar sólo en dos o tres reactivos más podría aumentar las calificaciones verbales y matemáticas hasta en 20 o 35 puntos. Sin embargo, en ge-

neral, la afirmación de The Princeton Review (Biemiller, 1986) y otras organizaciones de que las calificaciones de la SAT pueden aumentar en 100 o más puntos no tiene fundamento (Powers, 1993). La defectuosa metodología de investigación de muchos estudios sobre la asesoría produce resultados confusos y no concluyentes (Bond, 1989).

Se dice que la última versión de la SAT, SAT I, es menos susceptible de admitir asesoría que sus antecesoras debido al mayor énfasis puesto en la interpretación de largos pasajes. La omisión de la subprueba de antónimos, cuyas calificaciones pueden mejorarse por la simple memorización de palabras y cierto conocimiento de asociaciones de palabras, también ha disminuido los efectos de la asesoría. Se han conservado las analogías, el trabajo de completar frases y la interpretación de párrafos largos, tareas que no sólo requieren de conocimiento de palabras (vocabulario), sino también de habilidades de razonamiento que son más difíciles de mejorar mediante una asesoría rápida. Un análisis de los resultados de un estudio de más de cuatro mil examinandos que presentaron la prueba SAT en 1995-1996 indicó que los efectos de la asesoría en las calificaciones de la Prueba de Razonamiento son mucho menores de lo que afirman las principales compañías comerciales de preparación de pruebas (Powers y Rock, 1999).

Las calificaciones de los exámenes de admisión a la universidad suelen mejorar un poco con el desarrollo de los estudiantes y la familiaridad con las pruebas. En particular, tomar cursos académicos rigurosos y estudiar álgebra, geometría y significados de palabras justo antes de la prueba puede mejorar los resultados. Con respecto a los procedimientos para resolver la prueba, pasar por alto los reactivos difíciles y regresar a ellos después de terminar el resto de los reactivos de la sección, buscar respuestas "razonables" para los reactivos con extensos párrafos de lectura, adivinar respuestas en forma razonada y estrategias por el estilo no provocarán milagros, pero sí pueden mejorar en cierta medida las calificaciones (vea las recomendaciones para resolver pruebas en la página 49). De cualquier modo, además de comprobar que los reactivos nuevos no estén sesgados, el ETS los examina para investigar su susceptibilidad a la asesoría y descarta o modifica aquellos en los que puede mejorarse el desempeño mediante una instrucción o ejercicios de corto plazo (Swinton y Powers, 1985).

Diferencias en las calificaciones de la SAT

Las calificaciones de las pruebas no son números fijos, invariables; están sujetas a errores de medición y a diferencias genuinas en cuanto a habilidades y otras características personales. Los funcionarios escolares suelen estar alertas ante diferencias temporales y demográficas en las calificaciones de las pruebas, y con base en sus observaciones se decide intervenir en lo que respecta a la instrucción individual, las modificaciones de los programas y la distribución de los fondos públicos para la enseñanza. La reducción en las calificaciones de pruebas de habilidad y aprovechamiento despierta preocupación especial.

Cambios anuales en las calificaciones de la SAT. Durante la década de 1970, en todo Estados Unidos, fue cada vez más obvio que estaba disminuyendo la media en las calificaciones de la SAT y otras pruebas estandarizadas de habilidades cognoscitivas que se aplicaban a estudiantes de bachillerato. Aunque la media en las calificaciones de la SAT aumentó en la década de 1950 y principios de la de 1960, hacia finales de ésta y en la de 1970 ocurrieron reducciones considerables. La media en las calificaciones de matemáticas de la SAT no disminuyó tanto como en las pruebas verbales durante este periodo, pero la caída también fue significativa. Ocurrieron disminuciones para ambos sexos, para todos los grupos étnicos y para los estudiantes de mayor y menor capacidad. Se percibieron similares tendencias al declive en las calificaciones promedio de la prueba ACT, la Prueba Minnesota de Aptitud Académica, las Pruebas Iowa de Desarrollo Educativo y la Prueba Comprensiva de Habilidades Básicas.

Se han planteado varias explicaciones para entender la disminución en las calificaciones hacia finales de la década de 1960 y en la de 1970: menor atención, preocupación y supervisión de los padres hacia los hijos; falta de motivación de los alumnos por desempeñarse bien: demasiada televisión; una sociedad más permisiva; maestros que prestan menos atención a los estudiantes (Elam, 1978), y la simplificación de los libros escolares (Hayes, Wolfer y Wolfe, 1996). Otras explicaciones incluían a las drogas, el sexo, la falta de incentivos económicos para obtener una buena educación, y el espaciamiento de los hijos en las familias (vea Zajonc, 1986).

En una extensa revisión sobre la baja en las calificaciones de las pruebas de habilidad, un equipo de asesoría especial no logró encontrar ninguna evidencia de que la causa fuera una mayor dificultad de las pruebas (Austin y Garber, 1982). Aproximadamente la mitad de la reducción general entre 1963 y 1970 se consideró como un resultado de los cambios efectuados en la composición de la muestra de estudiantes que realizaron la prueba. Pero los cambios en la composición de género, raza-etnia, y posición socioeconómica de la muestra que se sometió a la SAT ya habían manifestado sus efectos para 1970. De acuerdo con el equipo asesor, las demás reducciones detectadas durante la década de 1970 se debieron a fuerzas sociales más constantes. No estaba claro exactamente cuáles fueron estas fuerzas y cuánta influencia tuvieron cada una, pero se mencionaron factores como programas de bachillerato menos exigentes intelectualmente, menores estándares educativos, maestros con habilidades inferiores, cambios en la estructura social y en los roles de las familias en Estados Unidos, la televisión, el desajuste nacional hacia principios de la década de 1970, y una menor motivación estudiantil. Más recientemente, Williams y Ceci (1997) observaron que el conjunto de estudiantes del último año de bachillerato que sustentó la prueba SAT fue menos selectivo en las citadas décadas de 1960 y 1970, y que aumentó la cantidad de instituciones demandantes de la prueba. Estos autores especularon sobre que si la SAT se hubiese aplicado a todos los estudiantes del último año de bachillerato y no a una muestra autoseleccionada en la década de 1950 y principios de la de 1960, la disminución en las calificaciones observadas hacia finales de la década de 1960 y principios de la de 1970 habría sido considerablemente menor (vea Berliner y Biddle, 1995).

Diferencias de género en las calificaciones de la SAT. A lo largo de los años, consistentemente, los hombres han superado a las mujeres en las calificaciones de la sección matemática de la SAT, pero hasta 1972 las mujeres tuvieron mejores calificaciones que los hombres en la parte verbal de esta prueba. En 1998, la media de la calificación era 37 puntos más elevada para los hombres que para las mujeres en la sección matemática y 7 puntos más elevada para los hombres en la parte verbal. En promedio, los hombres tuvieron calificaciones ligeramente más elevadas que las mujeres en la SAT. Sin embargo, estas fueron diferencias generales y no se presentaron las mismas para todos los grupos étnicos.

Durante mucho tiempo los críticos han argumentado que la SAT subestima los grados universitarios de las mujeres y, por lo tanto, que está sesgada en su contra (Shea, 1994). De acuerdo con Bob Schaeffer de FairTest (Chavez, 1993, p. A23):

> La misma naturaleza del SAT, que es una prueba de ritmo rápido, intensa presión y de opción múltiple con altas ventajas por adivinar, es un juego en el cual los chicos destacan. Quién sabe cuál es la razón cultural o biológica, pero las mujeres se inclinan más por intentar reflexionar sobre un problema, evalúan todas las opciones. Y eso las pone en desventaja estratégica.

Se afirmó que, como resultado de la brecha generacional en las pruebas SAT, las jóvenes tienen menos probabilidades que los hombres de obtener becas escolares. Los funcionarios del ETS contestaron que las diferencias entre las calificaciones SAT promedio entre hombres y mujeres reflejaban auténticas diferencias educativas y que la validez de predicción de la prueba es tan elevada para un sexo como para el otro. En cualquier caso, en la mayoría de los estados de la

Unión Americana no se otorgan becas universitarias con base en las calificaciones de la SAT únicamente, sino que se toman en cuenta otros criterios como el promedio de puntuación por grado y el desempeño en actividades extracurriculares.

Las causas de las diferencias de género en las pruebas SAT, que son las inversas a las diferencias en los promedios de la puntuación por grado en bachillerato y el primer año universitario, no están del todo claras. Las autoridades no están seguras de culpar a las pruebas, las escuelas, los factores biológicos o a otras variables ambientales. Otro factor posible es que, en promedio, la condición socioeconómica de las mujeres que presentaron la SAT en la década de 1980 era inferior a la de los hombres; y una hipótesis más establece que durante esta década las adolescentes estaban más preocupadas por sus citas románticas y el riesgo de embarazarse y se dedicaban menos al trabajo escolar que en la década de 1970 (Cordes, 1986). Cualesquiera que puedan ser las causas de las diferencias sexuales en las calificaciones de la SAT, al parecer están declinando: las mujeres han ido alcanzando a los hombres en ambas secciones de la SAT en años recientes (Shea, 1994).

Diferencias étnicas en las calificaciones de la SAT. Durante los años de 1990, las posiciones relativas de la población asiático-americana, afro-americana, mexicano-estadounidense, puertorriqueña y blanca en la SAT permanecían bastante constantes. A excepción de las calificaciones de los asiático-americanos en la sección matemática de la SAT, en 1998 las calificaciones promedio de los grupos minoritarios eran inferiores a las de los blancos en las secciones verbal y matemática de la prueba. En ese año, la media de las calificaciones verbal y de matemáticas de los negros era inferior en aproximadamente 100 puntos a la de los blancos. Los críticos sostenían que esta diferencia se debía al hecho de que la SAT estaba sesgada en contra de los negros. Pero ocurría algo similar con diferentes grupos étnicos en las pruebas de lectura, matemáticas y ciencia de la Evaluación Nacional del Progreso Educativo. Las calificaciones SAT inferiores para las minorías sin duda se deben, al menos en parte, a los más bajos ingresos familiares y niveles educativos de los padres. Sin importar las causas, las calificaciones de las pruebas de afro-americanos y latinos se han incrementado ligeramente en años recientes.

Estudiantes atletas y la SAT. Relacionado con, pero obviamente no exclusivo de, el problema de las diferencias de grupo étnico en las calificaciones de los exámenes de admisión está el requisito de la NCAA de que los estudiantes atletas tengan al menos un promedio de C y una calificación aprobatoria en la SAT para ser candidatos elegibles como estudiantes de primer grado en la escuela de la División I. Más precisamente, un estudiante que desee participar en competencias intercolegiales debe tener un promedio de puntuación de 2.5 o mayor y una calificación total en la SAT de 820 o más, un requisito que elimina a muchos estudiantes. Los adversarios de esos requisitos establecidos por la NCAA los han calificado de discriminatorios contra las minorías y sostienen que deberían reducirse. Pero aparentemente la mayoría de los representantes de las universidades de la División I de la NCAA consideran que los estudiantes atletas deberían ser capaces de cumplir con dichos requerimientos (Robbins y Almond, 1992).

OTROS TEMAS EN LAS PRUEBAS EDUCATIVAS

Aunque los asuntos relativos al SAT y a otros programas de evaluación nacionales han recibido más atención por parte de los medios de comunicación, otros aspectos relacionados con la evaluación en y por las escuelas también merecen tenerse en cuenta.

Trampas en las pruebas

Hacer trampa en las pruebas es un asunto preocupante en todos los niveles del sistema educativo. Al adquirir mayor importancia las calificaciones de las pruebas para determinar el futuro

educativo y las carreras profesionales de los individuos, pero además en la arena política para evaluar a las escuelas y otras instituciones, la tentación de hacer trampa parece haber aumentado. La administración de una prueba *segura* implica procedimientos estandarizados tales como verificar la identificación personal, sentar a los alumnos en determinada ubicación, una vigilancia cuidadosa y hojas de respuestas para disminuir las trampas, pero ninguno de estos procedimientos las elimina del todo. Las presiones de los padres, los maestros, los compañeros y los propios alumnos por tener buenos resultados pueden orillar a los estudiantes a robar pruebas, copiar respuestas de sus compañeros y hacer otro tipo de trampas.

Además de las observaciones directas de las trampas en las pruebas o los informes de otras personas acerca de las trampas que han realizado estudiantes específicos, pueden obtenerse pruebas de estas anomalías a partir de (1) patrones similares de respuestas erróneas idénticas de estudiantes que se sentaron juntos durante la prueba (Belleza y Belleza, 1989, 1995) y (2) gran cantidad de borraduras en la hoja de respuesta, sobre todo al cambiar respuestas erróneas por correctas. Esta última técnica se usó en California a mediados de la década de 1980 para confirmar las sospechas de que los aumentos drásticos en las calificaciones en algunas escuelas se debían a que los propios maestros cambiaban las respuestas de los alumnos en las pruebas del Programa de Evaluación de California (CAP). Las hojas de respuesta de las pruebas CAP, que medían habilidades básicas de lectura, escritura y matemáticas, se aplicaban anualmente en los grados tercero, sexto, octavo y doceavo de las escuelas públicas de California y se calificaban mediante escaneo electrónico de los datos. Las máquinas no sólo calificaban las hojas de respuesta, sino que también contaban las borraduras. Usando este procedimiento en combinación con trabajo de oficina para confirmar, se descubrió que en varias docenas de escuelas de Los Ángeles el porcentaje de borraduras era considerablemente mayor al esperado 3%. Aunque el furor resultante y la cobertura de la prensa al respecto precipitó fuertes protestas por parte del sindicato de maestros y el rechazo de algunos maestros a manejar las pruebas CAP, estos acontecimientos llevaron a investigaciones sobre trampas y alteraciones directas e indirectas en las pruebas CAP y CTBS (Banks, 1990).

Las alteraciones por parte de los maestros en las hojas de respuesta de los alumnos no pudieron refutarse en forma convincente, pero ¿por qué lo hicieron? La respuesta general parece ser que las calificaciones de las pruebas han llegado a ser tan usadas en la sociedad estadounidense —no sólo para evaluar a los individuos sino también a las escuelas, los distritos escolares, las etapas, e incluso los vecindarios—, que es enorme la presión sobre los maestros y las escuelas para que los alumnos se desempeñen bien. No sólo las presiones sociales sobre todas las personas relacionadas con las escuelas provocaron que los estudiantes hicieran trampa y los maestros alteraran las pruebas, sino que además éstos a menudo *enseñan para las pruebas*. Esta práctica y la alteración de pruebas es comprensible cuando consideramos la amplia publicidad que las escuelas dan a las calificaciones de prueba, la necesidad de justificar los aumentos en los gastos de educación y los incentivos por los cuales se asignan fondos adicionales a las escuelas cuando sus estudiantes obtienen calificaciones elevadas en las pruebas estandarizadas.

El círculo vicioso en que el superintendente estatal es presionado por los políticos, los superintendentes de distrito por el superintendente estatal, los directores de escuela por el superintendente de distrito, los maestros por los directores, los estudiantes por los maestros y los padres, y los políticos, directores y maestros por los padres, lleva a una situación en donde "siempre tienes a alguien encima de ti". Los directores y otros directivos escolares, que no tienen un puesto asegurado y pueden ser depuestos o transferidos si las calificaciones de los estudiantes resultan demasiado bajas en las pruebas estandarizadas, son particularmente susceptibles a la presión. Dado que sólo son humanos, es posible que dirijan esta presión a los maestros con el propósito de que sus escuelas den una buena impresión en el registro anual de promedios de calificaciones de pruebas de las escuelas, el cual se publica en los periódicos locales.

Los estudiantes, maestros y el personal administrativo de las escuelas requieren de alguna fuente de motivación para mejorar los bajos niveles en que la educación pública ha caído en muchas secciones escolares de Estados Unidos, así como algunas formas de evaluar la eficacia de sus esfuerzos. Sin embargo, la atmósfera de paranoia que según se informa permeó las posiciones de maestros y personal administrativo en el Distrito Escolar de Los Ángeles durante el escándalo de la alteración de pruebas de 1986 a 1988 no fue benéfica para la educación en general ni para la evaluación educativa en particular.

El efecto del lago Wobegon

En 1988 se informó que en Estados Unidos 70% de los estudiantes, 90% de los 15,000 distritos escolares, y los 50 estados tenían calificaciones superiores a las normas nacionales sobre las pruebas de aprovechamiento con referencia a normas aplicadas en escuelas elementales (Cannell, 1988). Este informe condujo a acuñar el término "efecto del Lago Wobegon", según la comunidad de ficción de Minnesota ideada por Garrison Keilor "donde todos los niños son superiores al promedio". Los hallazgos de Cannell se apoyaron en los resultados de un estudio realizado por el Departamento de Educación de Estados Unidos: 57% de los estudiantes de la escuela elemental tuvieron calificaciones superiores a la media nacional en lectura y 62% superiores a la media nacional en matemáticas. En otro estudio, llevado a cabo por los Amigos de la Educación, se descubrió que 83% de 5,143 distritos escolares, 73% de 4,501 distritos de escuelas secundarias, y todos menos dos estados (Louisiana y Arizona) estaban por encima del promedio en las calificaciones de pruebas de aprovechamiento (Cannell, 1989).

Una explicación para el efecto del Lago Wobegon es que se trata de una consecuencia de que las pruebas no recibían nuevas normas con la frecuencia necesaria. Otra explicación es que se debe a que los maestros asesoran a los alumnos en las preguntas de la prueba, y les permiten un tiempo mayor al establecido para responderlas, e incluso modifican las hojas de respuesta ya completadas.

Los editores de las pruebas de aprovechamiento estandarizadas citadas en estos estudios (CTB/McGraw-Hill, Riverside Publishing Company y Harcourt Brace) respondieron que resulta caro modificar las normas de las pruebas con la frecuencia que pudiera esperarse y que el aumento en las calificaciones de hecho puede indicar que las escuelas están mejorando. Sin embargo, los editores podrían esforzarse más por enfatizar ante los usuarios de las pruebas cuándo (fecha) y en qué muestras de estudiantes se estandarizaron sus pruebas. En particular, debería esclarecerse si se excluyó a algún grupo (por ejemplo, estudiantes de educación especial o aquellos con un dominio limitado del inglés) al seleccionar las muestras de estandarización.

Aunque la mayoría de los funcionarios no respondió por escrito o en forma impresa a los descubrimientos y críticas de Cannell, un experto en evaluación escolar afirmó que no es ético ni está garantizado suponer que ha habido trampa cuando aumentan las calificaciones. Este funcionario defendió el derecho de los maestros a examinar el contenido de una prueba a fin de determinar en qué áreas de habilidad necesitan mejorar los estudiantes, pero no a enseñar de acuerdo con la prueba (Landers, 1989).

Se reconoce ampliamente que las calificaciones tienden a dispararse hacia arriba cuando una batería de pruebas en particular se usa a lo largo de varios años en una escuela. Una razón del aumento puede ser que los maestros estén enseñando de acuerdo con la prueba, pero la explicación más plausible es que están enseñando *a partir* de la prueba (Lenke, 1988). Los maestros toman nota de las áreas de la prueba donde están bajas las calificaciones e intentan mejorar el conocimiento y las habilidades de los estudiantes en dichas áreas. Ésta es, desde luego, una estrategia de instrucción apropiada y no debe etiquetarse como trampa. También podríamos argumentar que el problema es con los tests con referencia a normas y que los resultados de las pruebas con referencia a criterios producirían información más significativa concerniente a las ventajas y deficiencias aca-

démicas y estarían menos sujetos a la mala interpretación. Como quiera que fuese, los políticos, los padres y otros interesados sin duda continuarán exigiendo datos de pruebas comparativos de un año al otro y entre escuelas para colaborar en la toma de decisiones educativas.

Pruebas y estándares educativos nacionales

La preocupación nacional de que los niños estadounidenses no están tan bien capacitados en ciencia y matemáticas como los niños de otros países data de por lo menos el lanzamiento del primer *Sputnik* soviético en 1957. Los resultados de pruebas de aprovechamiento aplicadas a nivel internacional reavivaron subsecuentemente esta preocupación al revelar que los escolares estadounidenses están atrasados con respecto a sus contrapartes de la mayoría de las demás naciones industrializadas en matemáticas y ciencia en particular (Centro Nacional para Estadísticas de Educación, noviembre de 1996, junio de 1997, febrero de 1998, 2001).

El Acta Nacional de Estándares de Habilidades, que se incorporó en el documento Metas 2000: Acta de 1994 para Educar a Estados Unidos, estableció un consejo de estándares de habilidad nacionales para desarrollar un sistema nacional voluntario de estándares, evaluaciones y certificaciones de habilidad. Esta ley exigía que se formularan diversos sistemas de evaluación no discriminatorios (evaluaciones orales y escritas, evaluaciones de portafolio, pruebas de desempeño, y otras por el estilo) y que se aplicaran para verificar el logro de estos estándares.

Se supuso que un conjunto de estándares educativos y las pruebas correspondientes proporcionarían una fuente de motivación y una guía para mejorar el aprendizaje en las escuelas públicas, así como una forma de determinar los progresos en la consecución de los estándares. Como se vio en el candente debate suscitado a finales de la década de 1990 acerca de la evaluación propuesta para toda la nación en el cuarto grado en lectura y en el octavo grado en matemáticas, ha sido difícil conseguir un apoyo bipartita para impulsar tales pruebas. Los conservadores tal vez teman que las pruebas nacionales sean el primer paso de la intromisión federal en las escuelas de sus vecindarios y que las escuelas locales estarían presionadas para adaptar sus planes de enseñanza con el fin de garantizar que los alumnos obtengan buenos resultados en las pruebas. Muchos representantes liberales se oponen a la evaluación nacional porque temen que las pruebas resulten discriminantes contra los niños de grupos minoritarios (Shogren, 1997).

En conexión con el Acta Nacional de Estándares de Habilidades, también ha habido una gran cantidad de debates entre los líderes gubernamentales y los profesionales en cuanto a la creación de pruebas nacionales de inglés, matemáticas, ciencia, historia y geografía para aplicarse a nivel nacional en los grados cuarto, octavo y doceavo. En diciembre de 2001 el Congreso de Estados Unidos aprobó un proyecto de ley que establece pruebas estatales anuales en lectura y matemáticas para todos los niños de los grados tercero al octavo, empezando desde el año escolar 2005-2006. Las escuelas donde las calificaciones no mejoren durante dos años consecutivos podrían recibir más ayuda federal. Si las calificaciones en dichas escuelas continúan sin elevarse, los estudiantes de bajos ingresos podrían ser candidatos a clases individuales o trasladarse a otra escuela pública con recursos federales. Si las calificaciones de una escuela aún no mejoran en cinco años consecutivos, el resultado podría ser cambios en el personal u otras consecuencias importantes, tales como la toma del mando por las autoridades estatales o la transformación del plantel en una escuela con exenciones. (*Los Angeles Times*, 9 de diciembre de 2001, p. A30.) Este proyecto de ley permite que distintos estados apliquen pruebas distintas, pero todos los estudiantes de un determinado estado tienen que presentar una prueba estatal para poder realizar comparaciones por grado, escuela y distrito, y un estado no está autorizado a cambiar de una prueba a otra cada año. Asimismo, las pruebas no sólo deben contener reactivos de opción múltiple, sino también preguntas abiertas que demanden a los estudiantes formular las respuestas y demostrar un razonamiento crítico.

Además de las evaluaciones de dominio en los grados escolares, se han realizado esfuerzos para obtener apoyo y desarrollar una prueba nacional que determine la medida en que los estudiantes universitarios han adquirido habilidades en razonamiento crítico, resolución de problemas y comunicación, las cuales son necesarias "para competir en una economía global y ejercer los derechos y responsabilidades de la ciudadanía" (Zook, 1993, p. A3). Las propuestas para que se realice una evaluación nacional de estudiantes posterior a la secundaria, que han sido estimuladas por la demanda de representatividad en la educación superior, también son controvertidas. Sin embargo, es posible que en el futuro cercano se desarrolle algún tipo de procedimiento evaluativo para determinar si las grandes sumas de dinero que se gastan en la educación superior son eficaces para equipar a los adultos jóvenes con las habilidades requeridas en el campo de trabajo. El desarrollo de tal prueba o pruebas sería caro, pero no resultaría tan costoso como tener un país lleno de graduados universitarios con una educación deficiente.

Evaluación de la inteligencia en las escuelas

Durante las últimas décadas, las relaciones entre experiencia educativa, estatus socioeconómico, etnia, nacionalidad, género, nutrición y muchas otras variables psicosociales y biológicas y las calificaciones obtenidas en pruebas de habilidades cognoscitivas se han considerado en cientos de investigaciones (vea el capítulo 8). Una pregunta constante se refiere al carácter de la interacción entre herencia y ambiente para determinar las calificaciones que se obtienen en las pruebas psicológicas. El significado de esta pregunta y sus implicaciones sociales y educativas han dado lugar a acciones legales en algunos estados. Están en tela de juicio algunas preguntas relativas a la utilidad y al sesgo de los tests de inteligencia. ¿Son estas pruebas útiles y justas para todos los grupos de niños, o están sesgadas en contra de ciertos grupos étnicos?

Entre los casos legales que han abordado la aplicación de pruebas de inteligencia en las escuelas están: *Stell* contra *el condado de Savannah-Chatham* (1963), *Hobson* contra *Hansen* (1967), *Diana* contra *el Consejo Estatal de Educación* (1970), *Guadalupe* contra *el Distrito de la Escuela Elemental Tempe* (1972), *Larry P*. contra *Riles* (1979), *PASE* contra *Hannon* (1980), y la *NAACP de Georgia* contra *el Estado de Georgia* (1985). En el caso de *Stell* contra *el condado de Savannah-Chatham* se tomó una decisión que después fue revocada por el Tribunal de Distrito de Apelación de Estados Unidos. La corte dictaminó en ese caso que, debido a que los CI de los niños negros eran inferiores a los de los niños blancos, exigir que ambos grupos se integraran en las mismas escuelas sería mutuamente desventajoso. En *Hobson* contra *Hansen*, la corte estableció que las pruebas de habilidad colectivas discriminan a los niños de grupos minoritarios y, por lo tanto, no podían utilizarse para asignar a los alumnos distintos cursos de habilidades. En *Diana* contra *el Consejo Estatal de Educación*, la corte dictaminó que no podían usarse procedimientos de evaluación tradicionales para ubicar a niños mexicano-estadounidenses en clases de niños con retraso mental susceptibles de ser educados, en California, y que debían tomarse medidas especiales (por ejemplo, asesoría bilingüe) para evaluar a los niños de grupos minoritarios. La decisión de la corte en *Guadalupe*... fue que se evaluara a los alumnos en su lengua principal y se eliminaran las partes injustas de la prueba. Asimismo, se estableció que las calificaciones de CI debían ser por lo menos dos desviaciones estándar menores a la media y que otros determinadores, tales como las medidas de comportamiento adaptativo, tendrían que incluirse al tomar decisiones sobre si los niños deberían clasificarse como retrasados mentales.

En su libro *Bias in Mental Testing* (*El sesgo en las evaluaciones mentales*), Arthur Jensen (1980) afirmó que ni las pruebas verbales de inteligencia ni las no verbales están sesgadas de manera significativa en contra de niños nacidos en Estados Unidos pero pertenecientes a grupos minoritarios. Jensen sostenía que las pruebas de inteligencia y de otras habilidades cognosciti-

vas tienen validez predictiva para todos los grupos étnicos y que no son responsables de las diferencias entre dichos grupos. Como se expresó en la decisión sobre *Larry P.* contra *Riles* (1979), el juez Robert Peckham de la Corte Federal de Distrito de San Francisco no estuvo de acuerdo con Jensen. Después de concluir que las pruebas de CI negaban igual protección legal a los cinco demandantes negros de una demanda de clase, el juez Peckham ordenó continuara su anterior prohibición de aplicar las pruebas de CI con propósitos de colocación de niños negros en la escuela pública de California para retrasados mentales susceptibles de ser educados. Así, se dictaminó que las pruebas de inteligencia administradas individualmente están sesgadas en contra de los negros, y que el Departamento de Educación Estatal de California no podía usar estas pruebas para emitir un diagnóstico educativo ni para la colocación de niños negros en las escuelas públicas. A esta decisión contribuyó el hecho de que una cantidad desproporcionada de niños negros habían sido asignados a clases de EMR, a las cuales el juez Peckham llamaba "educación sin salida". Por consiguiente, se estipuló que la proporción de niños negros en clases de EMR debería concordar con su proporción entre la población general de escolares. En 1986, el juez Peckham emitió de nuevo su prohibición del uso de pruebas de CI en las escuelas públicas de California, aun cuando se obtuviera el consentimiento de los padres. Sin embargo, la decisión de la corte en *Larry P.* no prohibía el uso de todas las pruebas de inteligencia en las escuelas públicas de California y dichas pruebas continuaron utilizándose para ciertos fines.

Menos de un año después de emitido el fallo de *Larry P.* contra *Riles,* otro juez federal, John F. Grady, tomó una decisión muy diferente en un caso similar de Illinois. En este caso, *PASE* (Parents in Action on Special Education) contra *Hannon* (1980), se decretó "que las pruebas WISC, WISC-R y Stanford-Binet, cuando se usan bajo los estatutos legales '[otros criterios] para determinar el programa educativo apropiado para un niño' (bajo la Ley Pública 94-142)... no discriminan en contra de niños negros" (p. 883). Como resultado, las pruebas de inteligencia continuaron administrándose con propósitos de ubicación en clases especiales en las escuelas públicas de Illinois y de muchos otros estados. De manera similar al fallo de *PASE* contra *Hannon*, la corte decidió en *la NAACP de Georgia* contra *el Estado de Georgia* (1985) que las pruebas de inteligencia no discriminan en contra de los niños negros. También contrariamente a las disposiciones del caso *Larry P.*, en la decisión de *Georgia*... se concluyó que la presencia de cantidades desproporcionadas de niños negros en clases de EMR no constituye una prueba de discriminación. Por último, en septiembre de 1992 el juez Peckham levantó la prohibición sobre las pruebas de inteligencia en las escuelas públicas de California bajo el argumento de que no era justo para los padres negros que deseaban sí fueran aplicadas para decidir la ubicación educativa de sus hijos con problemas de aprendizaje (Bredemeier, 1991). De hecho, esta disposición anuló la prohibición anterior (1986) en contra del uso de pruebas de inteligencia en las escuelas públicas de California. Una reseña de los casos citados y de otros presentados ante los tribunales, y que se relacionan con la evaluación de la inteligencia en las escuelas, revela que las decisiones judiciales han variado de un estado a otro y de acuerdo con el clima político de la época.

Aunque el uso de pruebas de inteligencia en ocasiones puede estimular la discriminación e incluso contribuir a una profecía que se cumple por sí misma, varios psicólogos y educadores sostienen que existen tres ventajas de usar estas pruebas con propósitos de ubicación. En muchos niños remitidos por los maestros con el señalamiento de que requieren educación especial se descubre que eso no es necesario cuando se les somete de nuevo a las pruebas. De hecho, si no se usaran las pruebas, probablemente se asignarían más niños de grupos minoritarios a las clases especiales. Incluso quienes están ubicados en dichas clases con base en calificaciones bajas en las pruebas a menudo aprovechan la educación especial al grado de que se mejora su CI, y ya no resultan candidatos para esos servicios. Por último, podría preguntarse qué sucede con los niños que sí requieren educación especial pero no son identificados por que no se les administran

pruebas de inteligencia. ¿Cuántos escolares se retrasan cada año porque no cuentan con la educación apropiada para sus habilidades al ser ubicados en clases generales?

PRUEBAS DE EMPLEO Y SESGO

Tan importante como los asuntos concernientes al uso de pruebas en escuelas y universidades es el aspecto de la justicia de estos instrumentos en cuanto a propósitos de selección de empleo, colocación y promoción. Como resultado de la creciente preocupación por los derechos civiles, la importancia del tema se incrementó cada vez más durante la década de 1960. Debido a que las pruebas de empleo se habían validado sobre todo en miembros de la cultura blanca dominante, era razonable preguntarse si tenían alguna validez para los negros y otras minorías. Tal fue la situación en el caso de *Myart* contra *Motorola* (1964), donde el asunto era si una prueba usada con fines de selección podría considerarse racialmente discriminatoria.

Legislación sobre la igualdad en las oportunidades de trabajo

El Acta de 1964 sobre Derechos Civiles (en Estados Unidos) surgió a raíz del caso Motorola y otras críticas de la evaluación psicológica. El Título VII de esta acta prohibía específicamente la discriminación con base en la raza, el color, el país de origen, el sexo o la religión.[1] Una disposición de la Suprema Corte sobre el Título VII ocurrió en el caso de *Griggs et al.,* contra *Duke Power Company* (1971), que se relacionaba con una demanda interpuesta contra la compañía Duke Power por empleados negros. La demanda se enfrentaba al requisito que antes había establecido Duke Power de presentar un diploma de bachillerato y a las nuevas políticas de promoción y contratación que requerían calificaciones mínimas predeterminadas en la Prueba de Personal Wonderlic y en la Prueba de Comprensión Mecánica Bennett. El presidente de la Suprema Corte, Warren Burger, quien escribió la opinión mayoritaria en ese caso, concluyó que "si no puede demostrarse que una práctica de empleo que funciona para excluir a los negros está relacionada [significativamente] con el desempeño en el trabajo, tal práctica está prohibida" (*Griggs et al.,* contra *Duke Power Company,* 1971, p. 60). Pero el juez Burger también señaló que:

> nada en el Acta [de Derechos Civiles] excluye el uso de procedimientos de evaluación o medición; obviamente son útiles. Lo que el Congreso ha prohibido es dar a estos dispositivos y mecanismos poder de control a menos que se demuestre que son una medida razonable del desempeño en el trabajo. El Congreso no ha ordenado que se prefiera a los menos calificados con prioridad frente a los más calificados simplemente por sus orígenes como minoría. Lejos de menospreciar las habilidades en el empleo como tales, el Congreso ha hecho de esas habilidades el factor preponderante, de modo que la raza, la religión, la nacionalidad y el sexo sean irrelevantes. (*Griggs et al.,* contra *Duke Power Company*, 1971, p. 11.)

La intención de la decisión de la Suprema Corte en el caso *Griggs et al.,* contra *Duke Power Company* fue solicitar que los empleadores demostraran que las habilidades medidas por sus pruebas de selección y demás procedimientos de contratación estaban relacionadas con el puesto. El efecto inmediato de la decisión era evaluar de nuevo, y en algunas situaciones descontinuar, ciertas pruebas de selección por parte de las empresas y las organizaciones industriales. Posteriormente, el Congreso concluyó que el Título VII del Acta de 1964 sobre Derechos Civiles no se había aplicado en forma adecuada y que continuaba la discriminación contra las minorías y las mujeres.

[1]También están relacionadas con las prácticas de empleo justo el Acta de 1967 sobre Discriminación por Edad en el Empleo (ADEA) y el Acta de 1990 sobre Estadounidenses con Discapacidades (ADA). La ADEA declara prohibida la discriminación contra los empleados o candidatos de 40 años o mayores en todos los aspectos del proceso de empleo. Con la ADA, a los individuos calificados con discapacidades deben otorgárseles iguales oportunidades en todos los aspectos del empleo.

Esta conclusión llevó a una revisión del Acta de Derechos Civiles, el Acta de 1972 sobre Iguales Oportunidades de Empleo. El Consejo Coordinador de Iguales Oportunidades de Empleo (EEOCC), que fue establecido por el Acta de Iguales Oportunidades de Empleo, preparó entonces un conjunto de normas denominado Lineamientos Uniformes para Procedimientos de Selección de Empleados. Estas normas describían los procedimientos a seguir por empleadores, organizaciones laborales y agencias de empleo, y exponían

> que cualquier procedimiento de selección que opere para descalificar o afectar de alguna otra manera adversa a los miembros de cualquier grupo racial, étnico o de sexo en mayor grado que a otro grupo, se ha validado de acuerdo con estos lineamientos, y que no están disponibles procedimientos alternativos de empleo con igual validez pero con un efecto menos adverso. (Comisión Estadounidense sobre Iguales Oportunidades de Empleo, 1973, p. 20.)

Los lineamientos establecen además que para ser juzgadas como una forma válida de predecir el desempeño, la prueba o combinación de pruebas normalmente deberán abarcar al menos la mitad de las habilidades medibles confiables y el conocimiento correspondiente al trabajo.

La ley concerniente al impacto desigual de las prácticas de empleo con respecto a ciertos grupos se amplió en tres casos subsecuentes: *Estados Unidos* contra *Georgia Power Company* (1973), *Albemarle Paper Co*. contra *Moody* (1975), y *Washington* contra *Davis* (1976). En el caso de *Albemarle Co*. contra *Moody*, tras descubrir que el programa de evaluación de la compañía era inadecuado, la corte sostuvo que, incluso si una prueba es válida pero afecta de manera adversa el empleo de ciertos grupos, la organización debería hacer todos los esfuerzos posibles para encontrar un dispositivo de seguridad menos sesgado. La definición legal de *impacto adverso* sigue la regla de los cuatro quintos, de acuerdo con la cual se considera que está presente una situación de impacto adverso si un grupo tiene una tasa de selección que es cuatro quintos (80%) menor que la del grupo con la mayor tasa de selección. Por ejemplo, si cien negros solicitan un empleo y se contrata a 60 blancos (el grupo mayor), entonces puede decirse que existe una situación de impacto adverso cuando menos de (4/5)60 = 48 negros también son contratados. Según los lineamientos del EEOCC, se requiere que los patrones adopten técnicas de selección con el menor impacto adverso. En *Washington* contra *Davis* (1976), el tribunal amplió el criterio al que deberían relacionarse las pruebas de selección para incluir el desempeño en programas de capacitación para el empleo.

Una revisión de 1978 de los lineamientos del EEOCC sobre la selección de empleados (Comisión Estadounidense de Iguales Oportunidades de Trabajo, 1978) no fue tan estricta como la versión original al requerir que los empleados realicen estudios de validez diferencial. Al igual que sus antecesores, los lineamientos revisados se diseñaron para exigir que los patrones justifiquen el uso de pruebas y otros procedimientos de selección que excluyan cantidades desproporcionadas de miembros de grupos minoritarios y mujeres. Los lineamientos describen tres métodos de validación en que pueden confiar los patrones: validez con relación a criterio, validez de contenido, y validez de constructo, pero no están claros en cuanto a qué tan grandes deberían ser los coeficientes de validez. Además, aunque los lineamientos revisados establecen que usar las pruebas es legítimo cuando las calificaciones están relacionadas con el desempeño en el trabajo, no especifican a qué se refieren con "criterios relacionados con el puesto".

La relación con el puesto es un concepto importante en este contexto, porque el uso de pruebas que tienen un impacto adverso se justifica en ocasiones con base en la afirmación de que están relacionadas con el puesto. La incapacidad de los lineamientos del EEOCC para esclarecer lo que significa "criterios relacionados con el puesto", y otros problemas similares, impulsó a muchas empresas y organizaciones de servicios a suspender por completo el uso de pruebas para la selección de empleos. Los lineamientos se consideran por muchas autoridades técnica-

mente obsoletos, y en muchos casos los estudios de validez requeridos son demasiado costosos y de valor cuestionable.

Queda claro que la implicación de los lineamientos del EEOCC era que los gerentes de personal necesitan llevar a cabo estudios de validación de todos sus procedimientos de selección, no sólo de las pruebas psicológicas, para determinar si están significativamente relacionados con el éxito en el trabajo. En *Watson* contra *Fort Worth Bank and Trust* (1988), el Tribunal estableció que los dispositivos subjetivos del empleo, tales como las entrevistas, pueden validarse y que los empleados pueden alegar impacto adverso como resultado de prácticas de promoción basadas en entrevistas. Por costoso que pueda ser, las entrevistas y otros métodos menos objetivos que las pruebas deben someterse al escrutinio mediante estudios de validez apropiados.

Otro interesante caso ventilado en los tribunales y relativo a las prácticas de empleo justas fue *Wards Cove Packing Company* contra *Antonio et al.,* (1989). Los demandantes en este caso fueron trabajadores filipinos y esquimales de enlatadoras de salmón en Alaska, quienes sostenían que la compañía los estaba excluyendo de puestos con mejor paga como la reparación de maquinaria. La decisión judicial en este caso es importante, porque cambió el peso de la prueba al empleado para que demostrara que no era válida ni confiable la prueba psicológica usada con propósitos de promoción. La preocupación sobre esta decisión, que invirtió el tema central del caso de *Griggs* contra *Duke Power* condujo al Acta de 1991 sobre Derechos Civiles. Esta acta confirmó los principios del Título VII del Acta de 1964 sobre Derechos Civiles, pero esclareció la situación de que el peso de la prueba recae en el patrón. Otra importante disposición del acta prescribió efectivamente el uso de calificaciones límite diferenciales por raza, género u origen étnico, lo que tuvo el efecto de desechar el sistema de cuotas vigente durante más de dos décadas.

Otras demandas legales relacionadas con la selección educativa y en el empleo se han ocupado de los efectos de la acción afirmativa o de cuotas al negar la admisión a la universidad a asiáticos y caucásicos estadounidenses que cuentan con la habilidad requerida. Aunque la corte ha apoyado los procedimientos de admisión o de selección que favorecen a los grupos con poca representación (por ejemplo, en *Estados Unidos* contra *la ciudad de Buffalo*, 1985), durante la década pasada fueron significativas las propuestas de prescindir de los requisitos de acción afirmativa ordenados legalmente en las escuelas y en el lugar de trabajo.

Imparcialidad en las pruebas

Como lo implican los lineamientos del EEOCC, las pruebas educativas y psicológicas estandarizadas en muestras de blancos son inaceptables para usarse en la selección de candidatos negros y de otros grupos minoritarios. Utilizar tales pruebas con grupos distintos a aquellos sobre los que se estandarizaron plantea el problema de la imparcialidad en las pruebas. El concepto de imparcialidad en la evaluación psicológica y educativa tiene un significado más estadístico que el supuesto por los lineamientos del EEOCC. El punto de vista tradicional en la medición psicológica es que la *imparcialidad* de una prueba para distintos grupos depende de si los candidatos con igual probabilidad de desenvolverse bien en un criterio de desempeño tienen las mismas posibilidades de ser seleccionados. De acuerdo con esta definición, incluso si la calificación media de un grupo es menor que la de otro, la prueba no necesariamente es parcial o injusta. Los negros y otras minorías de Estados Unidos pueden alcanzar calificaciones promedio más bajas que los blancos en las pruebas de empleo, pero esto no revela nada sobre la imparcialidad de las pruebas en el sentido técnico. Sin importar cualquier diferencia en las calificaciones promedio de las pruebas de dos grupos distintos, tradicionalmente se ha afirmado que una prueba de selección de empleo es imparcial si predice el mismo éxito en el trabajo para todos los grupos de candidatos.

Después de llamar la atención hacia una falla estadística en la definición tradicional (regresión equitativa) de imparcialidad de una prueba, Thorndike (1971) propuso una definición opcional. La definición del *índice constante* de Thorndike especifica que las calificaciones habilitantes de una prueba deberían establecerse de tal modo que se seleccionen distintos grupos de candidatos en proporción a la cantidad de cada grupo capaz de lograr un nivel aceptable en el criterio de desempeño. Por ejemplo, si 30% de todos los aspirantes blancos y 20% de todos los negros se juzgan capaces de desempeñarse bien en un trabajo determinado, entonces las calificaciones habilitantes en una prueba de selección deberían determinarse de tal manera que se contrate a 30% de los aspirantes blancos y 20% de los negros.

Otra definición de la imparcialidad en las pruebas fue sugerida por Cole (1973), quien propuso se establecieran calificaciones límite por separado para los dos o más grupos distintos de aspirantes, de modo que la probabilidad de selección sea la misma para candidatos potencialmente exitosos en cada grupo. Supóngase, por ejemplo, que dos grupos distintos están compuestos por 50 y 100 aspirantes respectivamente. Si se ha determinado con anticipación que 50% de todos los candidatos puede desempeñar el puesto en forma satisfactoria, entonces debería contratarse a $50\% \times 50 = 25$ aspirantes del primer grupo y $50\% \times 100 = 50$ candidatos del segundo grupo. Dunnette y Borman (1979) sugirieron un procedimiento de selección de cuotas similar. Sin embargo, en su propuesta el porcentaje de aspirantes por seleccionarse está definido de antemano; entonces se aplican ecuaciones de regresión separadas para cada grupo.

Los lineamientos del EEOCC revisados aceptan que la imparcialidad en las pruebas no es un concepto fijo y que los expertos pueden disentir en cuanto a su significado. Cualquiera que sea la definición que se prefiera, debería tomarse en cuenta la gravedad relativa de los errores de aceptar o rechazar aspirantes equivocadamente. Esto implica que la imparcialidad de una prueba es un asunto relativo, dependiendo de si se considera más grave rechazar a un aspirante que debería haber sido aceptado (*falso negativo*) o aceptar uno que fracasará (*falso positivo*). La conciencia social puede dictar que el primer error es más serio, mientras que las consideraciones de beneficio y seguridad indican que el segundo error es digno de mayor preocupación. Desde este punto de vista, el significado de imparcialidad es un asunto de política social, y no sólo de psicometría.

Incluso cuando una prueba se considera imparcial en su conjunto, es posible que algunos reactivos individuales resulten injustos o estén sesgados contra un grupo en particular. Por ejemplo, ciertos reactivos pueden presentar una visión estereotipada de los grupos minoritarios y las mujeres de acuerdo con la ocupación, la educación, la familia y la recreación de alguna forma (Tittle, 1984). Para identificar el sesgo en los reactivos y protegerse contra ellos, los editores de pruebas suelen realizar revisiones dictaminadoras para detectar los estereotipos y la familiaridad del contenido de las pruebas respecto a grupos particulares. También se han diseñado diversos procedimientos estadísticos para determinar la presencia de sesgos en los reactivos o el funcionamiento diferencial de reactivos (DIF). Entre estos procedimientos se encuentran los índices transformados de dificultad de reactivos, correlaciones biseriales para determinar las discriminaciones en los reactivos, las curvas características de reactivos, y variantes de chi cuadrada tales como la estadística Mantel-Haenzel (Cole y Moss, 1989; Scheuneman y Bleistein, 1989).

La construcción de curvas características de reactivos es una de las formas más descriptivas de detectar el sesgo en los reactivos. De acuerdo con este enfoque, un reactivo carece de sesgo si su curva característica es la misma para los grupos que se comparan. En otras palabras, los examinados con iguales habilidades, sin importar el grupo al que pertenezcan, tienen las mismas probabilidades de acertar en el reactivo. Se han llevado a cabo estudios experimentales donde el contenido de una prueba se varía para determinar si distintos grupos responden de manera diferente y estudios de análisis factorial para definir si las respuestas de distintos grupos producen

los mismos factores, y se han conducido investigaciones acerca del sesgo en las pruebas y los reactivos (Cole y Moss, 1989, Tittle; 1984).

En 1984 se llegó a una solución conciliadora en lo que respecta al problema del sesgo en los reactivos, cuando el Servicio de Evaluación Educativa aceptó un acuerdo fuera de tribunales en una demanda que acusaba de sesgo social a exámenes de una franquicia de seguros en Illinois. Según los términos del acuerdo se aceptó que al elaborar los exámenes del seguro el ETS emplearía primero reactivos en que negros y blancos obtuvieran calificaciones más similares. Este enfoque, conocido como *acuerdo de la Regla de Oro,* por el nombre de la compañía de seguros involucrada en la demanda, se usó después en otros estados. Como quiera que fuese, el acuerdo de la Regla de Oro posteriormente fue objeto de gran cantidad de debates y rechazo (*Educational Measurement,* 1987, *6*(2); Anrig, 1987; Denton, 1988).

RESUMEN

En Estados Unidos es práctica común pedir a los estudiantes que aprueben un examen de competencia mínima antes de otorgarles un diploma de bachillerato, y solicitar que los maestros pasen una prueba de habilidad profesional para ser contratados o confirmados en sus puestos. Algunas escuelas y universidades también han aplicado un método de valor agregado para la evaluación de cambios en conocimiento y habilidades durante los años anteriores a la graduación.

Los estudios han revelado que existen numerosas evaluaciones en las escuelas, pero que a menudo maestros, padres y los propios estudiantes carecen de suficiente información y capacitación como para interpretar los resultados de las pruebas en forma precisa. En años recientes ha adquirido impulso la evaluación de habilidades en estudiantes y maestros de bachillerato. Además de evaluar tanto a los estudiantes como a los maestros, las pruebas e instrumentos similares se usan para evaluar los programas educativos y determinar la efectividad de otros procedimientos y programas de intervención.

Durante muchas décadas se ha atacado el contenido y los usos de las pruebas estandarizadas de habilidades cognoscitivas. Las pruebas de opción múltiple en general, y los exámenes de admisión donde hay mucho en juego tales como la prueba SAT en particular, han sido muy criticadas por ser indicadores no válidos de lo que pretenden medir, por violar el derecho individual a la intimidad, por ser injustas tanto con los estudiantes privilegiados como con los de situación desventajosa, y por impulsar hábitos de estudio deficientes y prácticas sociales y económicas no éticas.

El interés mostrado en la legislación sobre veracidad en la evaluación fue indicativo de la exigencia de que la industria de la evaluación se vuelva más abierta y responsable hacia el público. También ha sido causa de preocupación con respecto a la evaluación de capacidades el declive anual de las calificaciones en la SAT y en otras pruebas de habilidad aplicadas nacionalmente, así como los efectos de la asesoría y de las diferencias de género y étnicas en las calificaciones de prueba.

La legislación y los litigios sobre derechos civiles y oportunidades de trabajo equitativas han dado origen a la reglamentación sobre el uso de pruebas en las empresas y la industria. Los lineamientos federales para los procedimientos de selección de empleados describen las características que deberán tener las pruebas y otras medidas a fin de considerarse técnicas aceptables y válidas para la selección y colocación de empleados. El problema de la imparcialidad en las pruebas para los grupos minoritarios y en desventaja condujo a nuevas definiciones de *imparcialidad*. Los asuntos legales y técnicos resultantes de la consideración de los conceptos de imparcialidad y predicción diferencial han alertado a los psicólogos profesionales, a los jefes de personal y al público en general sobre la necesidad de un uso más responsable de las pruebas y otros procedimientos de evaluación.

PREGUNTAS Y ACTIVIDADES

1. Haga una lista de los argumentos a favor y en contra de la evaluación de la aptitud de (a) estudiantes de bachillerato, (b)candidatos a maestros de escuela y (c) maestros con experiencia.

2. Discuta objeciones específicas contra las pruebas estandarizadas en general y contra las pruebas de opción múltiple en particular.

3. Describa las críticas de la SAT y las respuestas a estas críticas por parte del Consejo de Exámenes de Admisión a la Universidad y el Servicio de Evaluación Educativa.

4. ¿Por qué podría la legislación sobre veracidad en la evaluación propiciar que los maestros enseñen para la prueba?

5. Analice la legislación establecida por el Congreso de Estados Unidos y los fallos de la Suprema Corte con respecto a la evaluación en el trabajo, empezando con el Título VII del Acta de 1964 sobre Derechos Civiles.

6. La imparcialidad en una prueba de aprovechamiento se define como "la medida en que los reactivos de una prueba constituyen una muestra representativa de lo que saben los examinados", mientras que la imparcialidad en una prueba de habilidad se define como "la medida en que las calificaciones de una prueba son capaces de predecir igualmente el desempeño de criterios de distintos grupos". Sin embargo, Thorndike sostenía que las pruebas son justas si "las calificaciones aprobatorias [de las pruebas se] establecen en niveles que... califiquen a los candidatos de dos grupos en proporción con la fracción de los dos grupos que alcanza un criterio de desempeño específico". ¿Por qué existen distintas definiciones de imparcialidad en las pruebas, y qué implican tales definiciones?

7. Remítase a las 30 calificaciones aparejadas de la tabla A.2 en el apéndice A. Suponga que *X* es la calificación de una prueba de selección de empleo y *Y* la clasificación de desempeño en el trabajo. Suponga también que las 30 calificaciones se obtuvieron de un grupo mayoritario de aspirantes al puesto, mientras que las siguientes 20 calificaciones aparejadas corresponden a un grupo minoritario de candidatos.

X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
40	64	34	41	52	46	50	39
62	48	48	44	42	38	32	30
40	32	56	64	18	26	68	42
52	40	48	36	46	34	60	60
36	31	24	54	64	65	44	48

Ahora suponga que 50% de los candidatos del grupo mayoritario, 25% de los del grupo minoritario, y 40% de todos los aspirantes realizan el trabajo satisfactoriamente ($Y = 50$ o mayor). ¿Es justa la prueba de acuerdo con la definición tradicional de imparcialidad? ¿Según la definición de Thorndike? ¿Para la definición de Cole? ¿Cuáles son los porcentajes de falsos positivos y de falsos negativos de cada grupo, y cómo afectan la imparcialidad de la prueba?

8. Investigue acerca de escuelas y cursos de asesoría de pruebas, así como sobre los materiales de asesoría de pruebas publicados que estén disponibles en su área geográfica. Intente localizar a seis o más estudiantes que hayan pagado por recibir asesoría o preparación para la SAT, la GRE o cualquier otra prueba aplicada a nivel nacional. Pregúnteles si la asesoría les ayudó a mejorar sus calificaciones en la prueba. ¿Qué evidencias mencionaron para demostrar los efectos benéficos de tal asesoría?

CAPÍTULO DOCE

INTERESES VOCACIONALES

Las calificaciones de las pruebas de inteligencia y de habilidades especiales figuran entre las mejores formas de pronosticar el éxito educativo y ocupacional. Tales pruebas son medidas del *desempeño máximo*, en cuanto a que indican lo que una persona es capaz de lograr en condiciones óptimas. En general, los cuestionarios e inventarios de preferencias y otras variables afectivas contribuyen menos que las medidas cognoscitivas a pronosticar el éxito en la escuela y el trabajo, pero son muy útiles en la asesoría vocacional y educativa. Estas medidas de *desempeño típico* a menudo se suman en forma significativa a la información obtenida de medidas previas de habilidad y desempeño.

Una desventaja de los instrumentos de evaluación afectiva es que la mayoría no son tan objetivos y, por ende, tan confiables como las pruebas cognoscitivas. Incluso es objeto de debate que los cuestionarios, inventarios de informes autodirigidos y otros instrumentos afectivos de medición merecen el nombre de *pruebas*. No obstante, muchos instrumentos afectivos tienen una confiabilidad muy respetable, validez apreciable para ciertos propósitos y otras características de una buena prueba.

Tres variables afectivas que han recibido una gran cantidad de atención por parte de la investigación son los *intereses*, las *actitudes* y los *valores*. Las medidas de intereses son el tema del presente capítulo y las medidas de actitudes y valores se consideran en el capítulo 13. Los capítulos 14 a 18 completan nuestro estudio sobre las medidas afectivas con una revisión de varios tipos de procedimientos e instrumentos de evaluación de la personalidad.

FUNDAMENTOS DE LA MEDICIÓN DE LOS INTERESES

La información sobre los *intereses* de una persona, o las preferencias por ciertos tipos de actividades y objetos, pueden obtenerse de diversas maneras. El método más directo, simplemente preguntar a alguien qué le interesa, tiene sus escollos. Por ejemplo, las personas con frecuencia tienen poco conocimiento sobre sus intereses vocacionales o sobre lo que conllevan las ocupaciones en particular. Sin embargo, en ocasiones estos *intereses expresados* son mejores pronosticadores que la información obtenida en forma menos directa y no deben pasarse por alto en situaciones de consejería vocacional. Los resultados de una amplia investigación realizada por Flanagan, Tiedeman y Willis (1973) mostraron, por ejemplo, que varios grupos ocupacionales eran más dispares en sus intereses expresados que en sus habilidades cognoscitivas. Por ejemplo, estudiantes de ingeniería obtuvieron calificaciones mucho mayores al promedio en cuanto a intereses mecánico-técnicos e intereses en las ciencias físicas, mientras que estudiantes de leyes obtuvieron calificaciones más altas en cuanto a intereses por el servicio público (política), actividades literario-lingüísticas, negocios y ventas.

Otros métodos para determinar los intereses incluyen observaciones del comportamiento tales como la participación en diversas actividades, inferir los intereses de una persona a partir de su conocimiento de terminología especial u otra información sobre ocupaciones específicas, y aplicar uno de entre las docenas de inventarios de intereses disponibles.[1] Estos cuatro métodos de la medición del interés —pedir que se expresen los intereses, deducir los intereses a partir del comportamiento observado, inferir los intereses a partir del desempeño en pruebas de habilidades y determinar los intereses en inventarios de lápiz y papel— son aplicables a la evaluación de los grupos de intereses básicos descritos por Super y Crites (1962). Estos ocho grupos de interés son: científico, seguridad social, literario, material, sistemático, de contacto, expresión estética, e interpretación estética.

Historia y escenario actual

Empezando con el trabajo de E. L. Thorndike (1912) y otros, la investigación sobre intereses no se ha limitado a los contextos aplicados; se han realizado muchos estudios sobre los orígenes y la dinámica de los intereses. No obstante, los métodos estandarizados de medición de intereses se desarrollaron inicialmente con propósitos de asesoría y selección vocacional. James Miner tiene el crédito de haber llevado a cabo el primer intento sistemático por diseñar medidas de intereses vocacionales relacionadas con el criterio y validadas por contenido. Un cuestionario de intereses elaborado por Miner en 1915 fue el estímulo para celebrar un seminario histórico sobre medición de intereses en el Instituto Carnegie de Tecnología en 1919 y condujo a la construcción de inventarios de intereses vocacionales estandarizados. Uno de los participantes en el seminario fue E. K. Strong Jr., quien, impulsado por el éxito de uno de sus estudiantes de doctorado (K. Cowdery) al diferenciar entre ingenieros, abogados y médicos con base en sus intereses, amplió estos esfuerzos al iniciar un programa de investigación para diferenciar entre personas de muchas vocaciones distintas con base en sus intereses (vea Donnay, 1997). La investigación de Strong y sus alumnos condujo al desarrollo del Formulario de Intereses Vocacionales para Varones de Strong y de un instrumento paralelo para mujeres a finales de la década de 1920 y en la de 1930. Otros acontecimientos sobresalientes en la historia de la medición de intereses fueron la publicación en 1939 del Registro de Preferencias Vocacionales de Kuder y la investigación sobre medidas objetivas de intereses realizada por los psicólogos del Cuerpo de la Fuerza Aérea de Estados Unidos durante la Primera Guerra Mundial. Muchos inventarios de intereses se publicaron después de la guerra, pero las modificaciones de los instrumentos originales de Strong y de Kuder siguieron siendo las más populares.

En la actualidad se aplican inventarios de intereses por varias razones en diversos ambientes. Tradicionalmente, estos instrumentos se han usado sobre todo en contextos de asesoría ocupacional y educativa en los niveles de bachillerato, universidad y rehabilitación vocacional. También se han usado ampliamente en la investigación sobre diferencias individuales y de grupo, tanto en la investigación básica para determinar el carácter, los orígenes y efectos de los intereses como en la investigación aplicada con fines de asesoría, selección y colocación vocacional. Otras aplicaciones de los inventarios de intereses incluyen asistencia en la toma de decisiones sobre pasatiempos, a mediados de una carrera profesional, prerretiro y jubilación (Hansen y Campbell, 1985). Los asesores académicos y vocacionales y los investigadores psicológicos son, sin duda, los mayores grupos de usuarios de inventarios de intereses vocacionales, pero los consultores industria-

[1]Un indicador preliminar del interés por un objeto, persona o situación en particular puede obtenerse también mediante procedimientos fisiológicos como la medición pupilométrica (Hess, 1965) o la falométrica (Harris y Rice, 1996; Pithers y Laws, 1995).

les, los administradores de desarrollo de carreras y los practicantes de recursos humanos también los encuentran útiles.

Desarrollo de intereses

¿De dónde provienen los intereses? ¿Cómo se desarrollan y cambian con el tiempo? Los intereses vocacionales de los niños pequeños tienen, por lo general, un elemento de fantasía. Los niños fantasean sobre ser glamorosos, talentosos, heroicos o aventureros, pero tales ilusiones pueden tener poco que ver con sus habilidades o conocimiento sobre lo que conllevan las ocupaciones en particular. Normalmente, los niños evolucionan de una etapa de fantasía a otra de transición hacia finales de la niñez y principio de la adolescencia, y por último llegan a una etapa más realista en el desarrollo de los intereses vocacionales durante la adolescencia y la primera juventud.

Aunque los intereses vocacionales no se vuelven muy específicos, realistas ni estables durante el bachillerato y más adelante, la orientación general de los intereses de una persona puede notarse muy pronto en la vida. Los niños pequeños tienden a participar en actividades que consideran apropiadas y evitan las que consideran inadecuadas para sí mismos (Tyler, 1964). También hacen distinciones entre los papeles de las personas y los de la vida. De acuerdo con Anne Roe y sus coautores (Roe y Klos, 1969; Roe y Siegelman, 1964), los intereses vocacionales y, por ende, las elecciones de carrera provienen de los tipos de relaciones que los niños tienen con su familia. Un ambiente familiar cálido, de aceptación, tiende a crear una orientación hacia las "personas", mientras que una atmósfera fría, reservada, con mayor probabilidad origina una orientación hacia los "objetos" o las "cosas". Desde una perspectiva de aprendizaje social, los intereses se consideran como resultado de un refuerzo diferencial al participar en determinadas actividades, además de la imitación y los modelos de personas que son importantes para el individuo.

El papel de la herencia

El ambiente afecta, desde luego, los intereses en un grado considerable, pero los hallazgos de un estudio realizado por Grotevant, Scarr y Weinberg (1977) sugieren que los niños nacen con una predisposición hereditaria a interesarse por ciertas cosas. En este estudio de 114 familias biológicamente relacionadas, se descubrieron muchas correlaciones significativas entre las calificaciones de los niños y los padres en un inventario de intereses. En contraste, se hallaron pocas correlaciones de importancia entre los intereses de los padres y sus hijos adoptivos en 109 familias. Los niños biológicamente relacionados eran más similares en sus patrones de interés que los niños que no tenían ningún parentesco, y los intereses de parejas de niños del mismo sexo eran más similares que los de sexo opuesto. Los resultados de un estudio Minnesota muy difundido sobre gemelos idénticos criados en forma separada también indicaron que las correlaciones entre los intereses de gemelos idénticos son mayores que entre los intereses de otros pares familiares (Bouchard *et al.*, 1983; vea también Betsworth *et al.*, 1994, Maloney, Bouchard y Segal, 1991; Waller, Lykken y Tellegen, 1995). Debido a que los gemelos idénticos tienen idénticas herencias, se ha interpretado que estos descubrimientos demuestran la influencia de la herencia en los intereses. En general, las pruebas de estudios de comportamiento genéticos demuestran que los intereses vocacionales están influidos por la genética; aproximadamente, de 40 a 50% de la varianza en intereses vocacionales es atribuible a la varianza genética. De acuerdo con Lykken, Bouchard, McGue y Tellegen (1993), la influencia genética funciona mediante la interacción gen-ambiente, en cuanto a que las personas con determinada constitución genética están expues-

tas a experiencias y actividades particulares. Asimismo, es más probable que los intereses vocacionales sean consecuencia de influencias ambientales no compartidas —exclusivas del individuo—, más que de influencias ambientales compartidas con otras personas (Betsworth *et al.*, 1994; Maloney *et al.*, 1991).

Una creencia común es que el comportamiento de los padres ejerce más influencia que la herencia al moldear los intereses de los niños, pero Sandra Scarr y sus colegas concluyeron que lo que hacen los padres aparentemente tiene poco efecto en los intereses de los hijos (Grotevant, Scarr y Weinberg, 1977). Más que intentar forzar o guiar a los niños hacia ciertas áreas de interés, estos investigadores recomendaron a los padres que proporcionen a sus hijos una amplia variedad de experiencias y modelos. Los niños tendrán así una mejor oportunidad de desarrollar las predisposiciones o inclinaciones que posean naturalmente hacia actividades específicas.

Si se acepta que la gente tiende a interesarse por actividades que realiza bien y que la herencia desempeña un papel significativo en determinar las habilidades y el temperamento, es plausible que la herencia afecte los intereses indirectamente mediante las habilidades, el temperamento y la estructura física. Por ejemplo, una persona con una base genéticamente alta de nivel de actividad, pero con un nivel de inteligencia bajo, probablemente tendrá poco interés en convertirse en un físico teórico que dedica la mayor parte de su tiempo a reflexionar sobre problemas científicos. Por otra parte, una persona temperamentalmente activa y físicamente capaz puede mostrar mayor interés por convertirse en atleta profesional.

Estabilidad de intereses

Los patrones individuales de gustos y rechazos empiezan a desarrollarse mucho antes de que el individuo haya tenido experiencias con ocupaciones específicas. Estos primeros intereses son relativamente inestables, pero para cuando un niño llega al noveno grado, y casi con seguridad hacia el undécimo grado, sus preferencias por tipos específicos de actividades han quedado bastante bien determinadas. Los estudios longitudinales que abarcan dos o más décadas han demostrado que los intereses son sumamente estables hacia el final de la adolescencia (Hansen, 1988; Strong, 1955). Usando datos de archivo del Inventario de Intereses de Strong, Hansen (1988) encontró que los intereses tanto de hombres como de mujeres eran muy estables a lo largo de extensos periodos, hasta de 50 años. Por otra parte, los intereses de una persona pueden cambiar incluso en la edad adulta y debe tenerse especial cuidado al interpretar los resultados de inventarios de interés aplicados antes del noveno grado (Crite, 1969).

VALIDEZ DE LOS INVENTARIOS DE INTERESES

Debido a la importancia de la guía académica y vocacional, los inventarios de intereses comercialmente disponibles han sido casi tan populares como las pruebas de inteligencia general y de habilidades especiales. Sin embargo, en comparación con las mediciones cognoscitivas, los inventarios de intereses no pronostican con mucha precisión los grados escolares ni el desempeño ocupacional. En promedio, las calificaciones de los inventarios de intereses se correlacionan en alrededor de .20 a .30 con las notas escolares, mientras que las calificaciones de las pruebas de inteligencia general se correlacionan en alrededor de .50 con el mismo criterio. Las calificaciones de los inventarios de intereses contribuyen a pronosticar la selección ocupacional, la persistencia y la satisfacción, pero usualmente el éxito en el trabajo está más relacionado con la capacidad que con los intereses (Campbell y Hansen, 1981; Kuder, 1963). Como es más proba-

ble que las personas eviten las ocupaciones que les desagradan que se incorporen a ocupaciones que les gustan, en los inventarios de intereses las calificaciones bajas tienden a pronosticar más lo que una persona evita hacer de lo que las calificaciones altas indican lo que se inclina a hacer (Dolliver, Irvin y Bigley, 1972; Zytowski, 1976).

Simulación

Como también es cierto de las pruebas de habilidades, la validez de los inventarios de intereses al pronosticar la elección ocupacional se ve afectada por factores presentes al responder las pruebas y por características personales. El que sean o no mentiras intencionales, las respuestas a los inventarios de intereses pueden no indicar los verdaderos intereses de la gente. En los inventarios de intereses ciertamente puede fingirse. Bridgman y Hollenbeck (1961) descubrieron, por ejemplo, que al indicarles inventar sus respuestas, estudiantes universitarios llenaron un inventario de intereses (la Forma D de Kuder) de tal modo que sus respuestas fueron muy similares a los de personas empleadas en ocupaciones específicas.

Simplemente porque los inventarios de intereses pueden simularse no necesariamente significa que eso ocurrirá. Estos instrumentos son menos útiles cuando es desventajoso dar informes falsos, lo que es más probable cuando se usan las calificaciones para propósitos de selección educativa o laboral. Sin embargo, responder con falsedad a un inventario de intereses es mucho menos probable cuando se aplica con fines de consejería académica y vocacional. Incluso cuando la gente al parecer pudiera beneficiarse de dar respuestas falsas en un inventario de intereses, no siempre lo hace. Por ejemplo, el Inventario de Intereses Vocacionales de Strong (SVIB) se usó durante muchos años en seleccionar a individuos para capacitación avanzada en la marina estadounidense. En tales circunstancias podría suponerse que la simulación sería un problema. Sin embargo, esto no resultó ser el caso (Abrahams, Neumann y Gilthens, 1971). Las calificaciones promedio de un grupo de jóvenes que respondió el SVIB como parte de la solicitud de beca para la marina fueron muy similares a las que obtuvieron en bachillerato un año antes o en la universidad un año después de solicitar una beca. Además, las correlaciones entre los perfiles de calificación de interés obtenidas en la situación de solicitud de beca y las obtenidas en condiciones de evaluación de rutina estuvieron en el .90. Podría haber sido ventajoso para los aspirantes producir un resultado más favorable mintiendo, pero al parecer no lo hicieron de ninguna manera perceptible.

Grupos de respuesta

Aunque no es lo mismo que la simulación intencional, la tendencia a responder a la estructura en lugar de al contenido de los reactivos de prueba (*grupos de respuesta*) también pueden dar como resultado calificaciones imprecisas en los inventarios de intereses. De particular preocupación son los grupos de respuesta de *aceptación* o de acuerdo más que de disentimiento cuando no hay certeza, y de *conveniencia social* o dar una respuesta socialmente más conveniente. Una técnica diseñada para reducir estos grupos de respuesta es el formato de *elección forzosa*. Los reactivos con este formato consisten en dos o más afirmaciones descriptivas que son iguales en cuanto a conveniencia social, pero distintas en contenido y validez. En un reactivo de intereses de elección forzada, se pide a los examinados indicar cuál de las actividades descritas en tres o cuatro opciones les gustaría más (M) hacer y cuál les gustaría menos (m) hacer (vea la figura 12.1). Desafortunadamente, en ocasiones a las personas les parece raro y frustrante el formato de elección forzada.

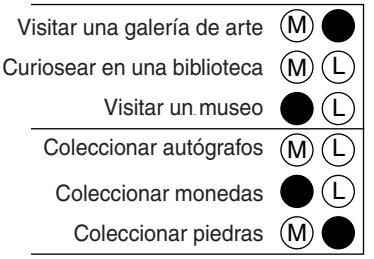


FIGURA 12.1 Muestra de reactivos del Estudio de Intereses Generales de Kuder.

(Tomada del Estudio de Intereses Generales de Kuder, Forma E, Hoja de respuestas. Reproducida con autorización del editor, National Career Assessment Services, Inc.® Todos los derechos reservados.)

Estatus socioeconómico

Un factor demográfico que está significativamente relacionado con las respuestas a los inventarios de interés vocacional, y por lo tanto con su validez, es la situación socioeconómica de quien responde. Las personas de clase trabajadora no siempre tienen la oportunidad de cultivar sus intereses o capacitarse y participar en ocupaciones que sean atractivas para ellas. Para estos individuos, la seguridad económica es un factor más importante en las decisiones sobre el empleo que satisfacer sus intereses. Ésta es una de las razones por las que, durante muchos años, los psicólogos mostraron poca inclinación a construir inventarios para medir los intereses vocacionales de las personas que planeaban incorporarse a ocupaciones que no requerían capacitación o que requerían capacitación parcial o incluso total. Como el dinero parecía ser un determinante ocupacional más importante que satisfacer intereses vocacionales, el desarrollo de inventarios de interés para ocupaciones no profesionales se consideró improductivo. Como consecuencia, los primeros inventarios de interés se diseñaron casi por completo para usarse en la asesoría de jóvenes que estaban planeando incorporarse a una profesión. La situación cambió en cierta medida después de la Segunda Guerra Mundial, pero el principal foco de los inventarios de interés permaneció en las profesiones.

En el extremo superior de la escala socioeconómica están los niños de familias adineradas. Ellos pueden tener fuertes intereses vocacionales, pero las expectativas y tradiciones familiares y sociales con frecuencia son más importantes que los intereses de los individuos para determinar las decisiones concernientes a sus carreras. Los hijos de familias acaudaladas pueden no estar autorizados para hacer lo que quieran, ya sea debido a que el estatus o la remuneración económica de las ocupaciones en que se interesan no son lo bastante altos o porque los padres esperan que sus hijos sigan sus pasos o hasta superen sus logros. Por otra parte, los jóvenes de clase media en ascenso tienen más probabilidades de intentar mejorar sus oportunidades de éxito ingresando en las ocupaciones donde tienen fuertes intereses, tal vez aunque no posean las habilidades que se requieren. Por esta razón, los inventarios de intereses en general han pronosticado mejor las elecciones ocupacionales para las personas de clase media que para las de clase alta o clase trabajadora (McArthur y Stevens, 1955). En cualquier caso, muchas de las ocupaciones del actual campo de trabajo no satisfacen los intereses de las personas que las realizan (Warnath, 1975). Así, ¿qué hacen las personas cuando descubren grandes discrepancias entre lo que les gustaría hacer y lo que deben hacer a fin de sobrevivir? En la mayoría de los casos, más que

arriesgar su seguridad en una inexorable búsqueda en pos de sus intereses y aspiraciones vocacionales, es mucho más probable que adapten sus aspiraciones para acercarse más a lo que de hecho les es posible alcanzar (Gottfredson y Becker, 1981).

INVENTARIOS DE INTERESES DE STRONG

Dos de los primeros y más notables inventarios para medir los intereses vocacionales fueron diseñados por E. K. Strong, Jr. y G. F. Kuder. Como resultado de una investigación realizada durante la década de 1920, Strong descubrió significativas diferencias consistentes en cuanto a los informes sobre sí mismos de lo que a las personas les gustaba o disgustaba. Decidió diseñar un inventario para evaluar las diferencias individuales en intereses, empezando con la elaboración de una variedad de reactivos referentes a las preferencias por ocupaciones específicas, materias escolares, diversiones, actividades y tipos de personas. Estos reactivos, además de una escala para clasificar las habilidades y características individuales, se aplicaron entonces a grupos de hombres empleados en ocupaciones específicas. Al comparar las respuestas de los sujetos ubicados por grupos ocupacionales con las de los hombres en general, Strong pudo desarrollar varias docenas de escalas ocupacionales consistentes en reactivos que una cantidad considerable de hombres de ocupaciones específicas respondió de manera distinta a la de los hombres en general. Este Formulario de Intereses Vocacionales para Varones de Strong fue la primera medición de intereses estandarizada y distribuida comercialmente. Varios años más tarde, cuando quedó claro que los intereses de las mujeres no se limitaban al trabajo de oficina, a la enseñanza elemental, a la enfermería y a las labores domésticas, se diseñó un instrumento paralelo, el Formulario de Intereses Vocacionales para Mujeres de Strong.

Por varias razones, incluyendo el deseo de acatar el Título IX del Acta de 1964 sobre Derechos Civiles (estadounidense) y de refutar las acusaciones de sexismo, las formas para hombres y mujeres del Formulario de Intereses Vocacionales®[2] se combinaron en 1974 en un solo instrumento, el Inventario de Intereses de Strong-Campbell (SCII). Se realizaron esfuerzos por eliminar el sesgo hacia el sexo en el contenido de los reactivos y las etiquetas ocupacionales y por crear un inventario más independiente del género. Sin embargo, se reconoció que el sesgo hacia el sexo se había reducido aunque no eliminado del todo del SCII.

Formato del Inventario de Intereses de Strong

La última edición del instrumento originado por Strong es el Inventario de Intereses de Strong® (SII) (CPP).[3] Este inventario consta de 317 reactivos agrupados en las siguientes ocho partes:

I. *Ocupaciones*. Cada uno de los 135 títulos se responde con (A) agrado, (I) indiferencia o (D) desagrado.

II. *Materias escolares*. Cada una de las 39 materias escolares se responde con (A) agrado, (I) indiferencia o (D) desagrado.

III. *Actividades*. Cada una de las 46 actividades ocupacionales generales se responde con (A) agrado (I) indiferencia o (D) desagrado.

[2]*Formulario de Intereses Vocacionales de Strong* y *SVIB* son marcas registradas propiedad de la Imprenta de la Universidad de Stanford.

[3]*Inventario de Intereses de Strong* y *SII* son marcas registradas propiedad de la Imprenta de la Universidad de Stanford.

IV. *Pasatiempos*. Cada una de las 29 diversiones o pasatiempos se responde con (A) agrado, (I) indiferencia o (D) desagrado.

V. *Tipos de personas*. Cada uno de los 20 tipos de personas se responde con (A) agrado, (I) indiferencia o (D) desagrado.

VI. *Preferencias entre dos actividades*. Para cada uno de los 30 pares de actividades se indica la preferencia por la actividad de la izquierda (I), por la de la derecha (D) o por ninguna de las dos (=).

VII. *Sus características*. Cada una de las 12 características personales se responde con Sí, ? o No, dependiendo de si describen o no a la persona.

VIII. *Preferencia en el mundo del trabajo*. Para cada uno de los 6 pares de ideas, datos y cosas, se indica la preferencia por el reactivo de la izquierda (I), por el de la derecha (D) o por ninguno de los dos (=).

Aunque los reactivos, el formato y el procedimiento de aplicación del Inventario de Intereses de Strong permanecieron básicamente sin cambios con respecto a la edición anterior, el perfil se amplió para incluir 211 escalas ocupacionales (102 pares con distintas escalas para hombres y mujeres y 7 escalas para ocupaciones representadas por un solo género).

Calificación

El SII se califica sólo por computadora y los procedimientos de ponderación y calificación de reactivos son un secreto comercial. Los inventarios resueltos se envían a Consulting Psychologist Press para su calificación, elaboración de perfiles e interpretación, o bien pueden calificarse e interpretarse por medio de programas de cómputo que se venden a los usuarios.[4] El informe del Perfil Strong muestra las calificaciones del examinando en cientos de escalas; también están disponibles otros tipos de informes, tales como el Informe Interpretativo de Strong, que proporciona información gráfica detallada sobre los intereses ocupacionales del examinando y descripciones a la medida sobre las mejores ocupaciones a elegir.

El SII se califica en cinco grupos de mediciones: Índices administrativos, Temas ocupacionales generales, Escalas de intereses básicos, Escalas ocupacionales y Escalas de estilo personal. Antes de intentar interpretar las calificaciones de una persona en las últimas cuatro categorías, es preciso verificar las calificaciones de tres índices: el Índice de respuestas totales; los Índices de porcentaje de "Agrado", "Indiferencia" y "Desagrado"; y el Índice de respuestas no frecuentes. El Índice de respuestas totales no debe caer por debajo de 300 (de entre 317); los Índices de Porcentaje de "Agrado", "Indiferencia" y "Desagrado" no deberán quedar fuera del rango de 14 a 60; y el Índice de respuestas no frecuentes no debe ser menor que cero (Harmon, Hansen, Borgen y Hammer, 1994). Los Índices administrativos aparecen en la parte inferior de la página 6 de las seis páginas de "Instantánea: un resumen de resultados".

Como se muestra en la figura 12.2, el SII se califica en seis temas ocupacionales generales. Estos seis temas, que se describen en la tabla 12.1, se basan en las seis categorías de las "personalidades vocacionales" de J. L. Holland (1985): Realista (R), Investigadora (I), Artística (A), Social (S), Emprendedora (E) y Convencional (C). El estándar del examinado de la calificación

[4]El SII también está disponible mediante Internet para personas calificadas que ya tienen contratada una cuenta con Consulting Psychologist Press.

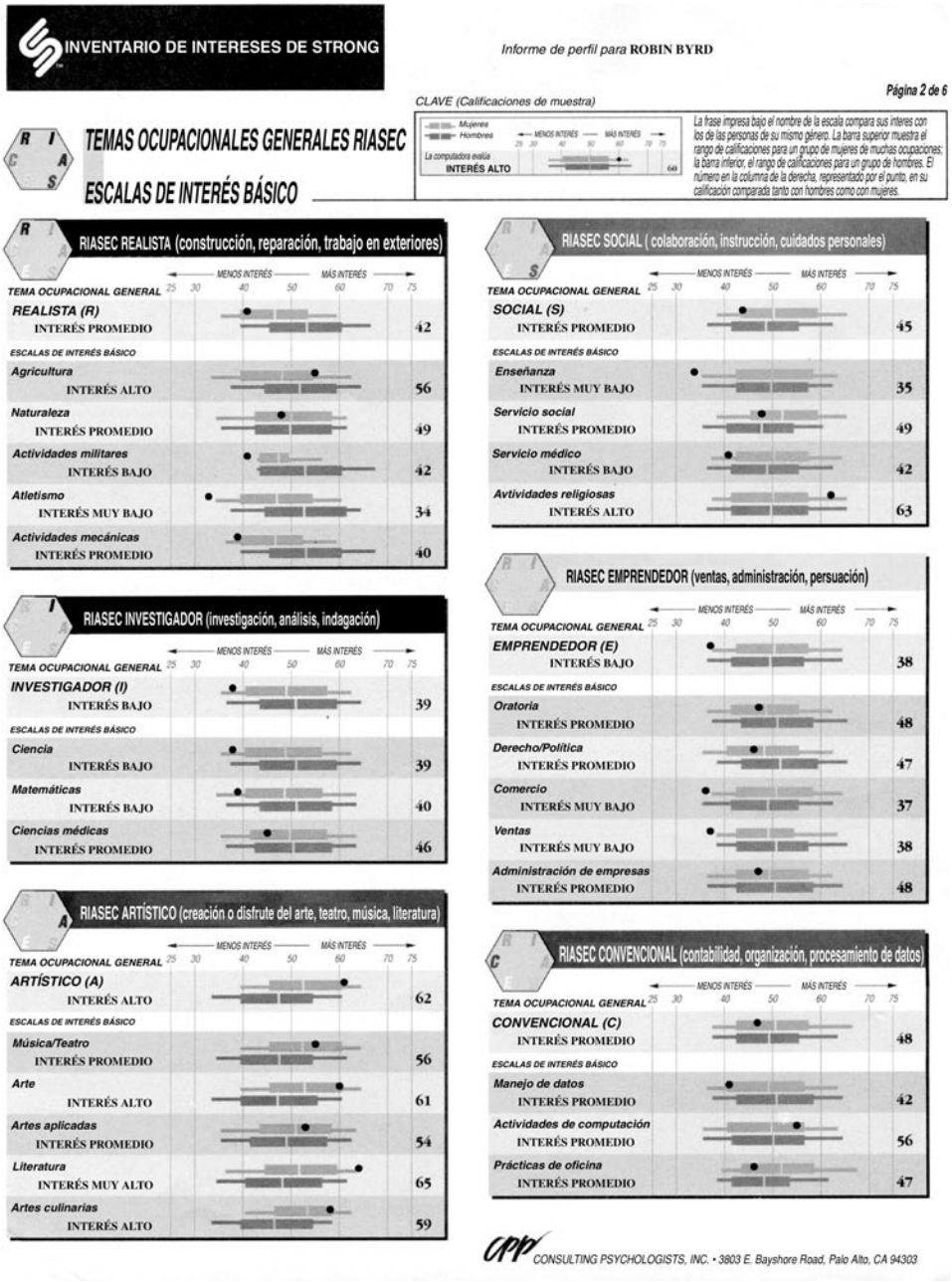


FIGURA 12.2 Página 2 de las calificaciones de muestra del perfil Instantáneo en el Inventario de Intereses de Strong.

(Modificado y reproducido con permiso especial del editor, Consulting Psychologist Press, Inc., Palo Alto, California 94303, a partir de *Aplicaciones y Guía Técnica del Inventario de Intereses de Strong* del Inventario de Intereses de Strong de los Formularios de Intereses Vocacionales de Strong, Forma T31.7. Lenore W. Harmon, Jo-Ida C. Hansen, Fred H. Borgen y Allen L. Hammer. Derechos reservados 1933, 1938, 1945, 1946, 1966, 1968, 1974, 1981, 1985, 1994, por el Consejo de Fiduciarios de la Leland Stanford Junior University. Todos los derechos reservados. Impreso con autorización de la Imprenta de la Universidad de Stanford, Stanford, California 94305.)

TABLA 12.1 Descripciones de los tipos RIASEC

Realista. Las personas de este tipo gustan de manipular herramientas, máquinas y otros objetos, estar al aire libre y trabajar con plantas y animales; les disgustan las actividades educativas y terapéuticas. Tienden a poseer aptitudes atléticas, mecánicas y técnicas, pero tienen deficiencias en habilidades sociales y educativas. Se caracterizan por ser personas prácticas, adaptables y naturales que probablemente eviten situaciones donde se requieran habilidades verbales e interpersonales y busquen trabajos como mecánico automotriz, granjero o electricista.

Investigador. Las personas de este tipo prefieren actividades que demanden una gran cantidad de reflexión y comprensión, por ejemplo, proyectos científicos. Les gusta observar, aprender, investigar, analizar, evaluar, resolver problemas, pero tienden a evitar situaciones que requieran habilidades interpersonales y de persuasión. Estas personas se caracterizan como racionales, cautelosas, curiosas, independientes, introvertidas. Es más probable encontrarlas en áreas como la química, la física, la biología, la geología y otras ciencias.

Artístico. Las personas de este tipo son imaginativas, introspectivas, complicadas, emocionales, expresivas, impulsivas, rebeldes y desordenadas. Prefieren situaciones no estructuradas donde la creatividad y la imaginación puedan expresarse. Es probable que les atraigan profesiones como las de actor, músico o escritor.

Social. Las personas de este tipo tienen habilidades verbales e interpersonales. Les gusta trabajar con la gente en su desarrollo, instrucción, información, capacitación, curación, ayuda o apoyo de todo tipo. Tienden a ser humanistas, empáticas, y cuentan con buenas habilidades para la enseñanza, pero mínimas habilidades mecánicas y científicas. Los demás las consideran cooperativas, amistosas, colaboradoras, convincentes, cautelosas y comprensivas. Suelen encontrarse en campos como la psicología clínica o consultiva, la terapia del lenguaje, la enseñanza y otros similares.

Emprendedor. Las personas de este tipo se describen como agresivas, ambiciosas, activas, dominantes, hedonistas, seguras de sí mismas y sociables. Tienden a tener grandes capacidades verbales y de liderazgo; les gusta influir y convencer a los demás, así como guiar o dirigir empresas por la recompensa económica. Prefieren trabajos como los de gerente, ejecutivo de negocios o vendedor.

Convencional. Las personas de este tipo se describen como conscientes, eficientes, inflexibles, obedientes, ordenadas, persistentes y autocontroladas. Tienden a tener buenas habilidades aritméticas y de trabajo de oficina y son adeptos a llevar archivos, realizar informes y procesar datos. Estas habilidades son congruentes con énfasis en el logro económico y empresarial que los conduce a trabajos como banquero, librero y experto en impuestos.

T y su posición de calificación en el rango medio de 50% de los grupos de norma de hombres y mujeres, en cada tema de RIASEC y de 3 a 5 escalas de intereses básicos que quedan por abajo, se incluyen en la columna situada al extremo derecho del recuadro correspondiente de la página 2 del resumen. Las 25 escalas de intereses básicos se construyeron agrupando reactivos con altas intercorrelaciones. Las calificaciones de estas escalas representan la fuerza y consistencia de áreas de interés especiales (Agricultura, Ciencia, Música-Teatro, Enseñanza, Oratoria, Manejo de datos, etc.). Las calificaciones *T* de las escalas se diseñaron como Muy poco interés, Poco interés, Interés medio, Interés alto o Interés muy alto.

Las calificaciones *T* en las 109 escalas ocupacionales, que consisten en calificaciones separadas para hombres y mujeres en 102 escalas y calificaciones de género combinadas en 7 escalas, se enlistan en el tema correspondiente a RIASEC de las páginas 3 a 5 de “Instantánea”. Cada

una de las escalas ocupacionales se construyó comparando las respuestas de hombres y mujeres empleados en una ocupación particular con las respuestas de un grupo de referencia de hombres y mujeres en general. Todas menos seis de las escalas ocupacionales se compararon con las escalas del género opuesto y se estandarizaron en forma separada por género. La puntuación cruda de una persona en una escala ocupacional dada se determina sumando los valores numéricos asignados a sus respuestas en la escala. El valor asignado depende de la dirección en que el reactivo discrimina entre hombres o mujeres empleados en dicha ocupación y hombres o mujeres en general. Una vez que se han sumado todos los valores correspondientes a las respuestas de una persona a los reactivos de una escala en particular, la calificación cruda obtenida se convierte a una calificación *T* que varía directamente de acuerdo con el grado de similitud entre las respuestas del examinado y las del grupo de la ocupación específica. Las calificaciones *T* se agrupan en tres categorías: intereses dispares, rango medio e intereses similares.

Los perfiles de las calificaciones *T* y los rangos medios de 50% de las calificaciones en los grupos de norma de hombres y mujeres en las cuatro escalas de estilo personal se proporcionan en la página 6 de la "Instantánea". Éstas son escalas bipolares con las siguientes definiciones:

Estilo de trabajo: "Trabaja con ideas/datos/cosas" contra "Trabaja con personas".
Ambiente de aprendizaje: "Práctico" contra "Académico".
Estilo de liderazgo: "Dirige con el ejemplo" contra "Dirige a los demás".
Toma de riesgos/Aventura: "Actúa con cautela" contra "Acepta riesgos".

Las calificaciones de las escalas de estilo personal también son útiles en la consejería vocacional y la exploración de carreras.

Características psicométricas

La muestra de desarrollo del Inventario de Intereses de Strong consistió en más de 55,000 personas en 50 ocupaciones que realizaron el inventario en 1992-1993. Sin embargo, sólo 9467 mujeres y 9,484 hombres de este grupo se usaron como muestras de referencia general. Esto bastó para permitir una validación precisa de las antiguas escalas ocupacionales y el desarrollo de otras nuevas. Con respecto a la confiabilidad del SII, la elevada consistencia interna y coeficientes de calificaciones de test-retest en los Temas ocupacionales generales, las Escalas de intereses básicos, Escalas ocupacionales y Escalas de estilo personal se han obtenido en grupos de estudiantes universitarios y adultos empleados. Los coeficientes alfa de Cronbach para los Temas ocupacionales generales abarcan de .90 a .94 en la muestra de referencia de hombres y mujeres, y los coeficientes test-retest para un intervalo de 3 a 6 meses en una muestra de 65 empleadas y 75 empleados van de .84 a .92. Los coeficientes alfa para las Escalas de intereses básicos abarcan de .74 a .94 en la muestra de referencia, y el coeficiente test-retest en una muestra de aproximadamente 200 hombres y mujeres abarca de .82 a .94. Para las escalas de estilo personal, los rangos alfa van de .78 a .91 en la muestra de referencia; el coeficiente test-retest (intervalo de 1 a 4 meses) en una muestra de 128 mujeres y 103 hombres va de .81 a .92. La gran mayoría de los coeficientes test-retest en las muestras de los estudiantes universitarios y los adultos empleados para las Escalas ocupacionales también estuvieron entre los .80 y .90.

Los diversos tipos de evidencia que corresponden a la validez de contenido, concurrente, predictiva y de constructo del SII se registran en el manual (Harmon, Hansen, Borgen y Hammer, 1994). La validez de contenido es quizá el tipo más fácil de establecer: un análisis de la composición del SII de 1994 apoya la validez de contenido de este instrumento. Las calificacio-

nes de clasificación media de los Temas ocupacionales generales, las Escalas de intereses básicos, las Escalas de estilo personal y las Escalas ocupacionales de los 109 grupos ocupacionales proporcionan evidencia de la validez concurrente y de constructo, así como las correlaciones entre mediciones similares en otros inventarios de intereses (por ejemplo, el Inventario de Intereses Vocacionales de Holland). De manera similar, las diferencias en las calificaciones medias de las Escalas de estilo personal por nivel educativo, especialidad educativa y ocupación refuerzan la evidencia de validez para estas escalas. Aunque todavía no se han recopilado muchos datos sobre la validez a largo plazo de este inventario, los datos de estudios de validez predictiva realizados con ediciones previas de la Escala Strong (Campbell, 1971; Hansen y Campbell, 1985) se han extrapolado a la edición de 1994.

INVENTARIOS DE INTERESES DE KUDER

En contraste con el variado formato de reactivos del Inventario de Intereses de Strong, G. F. Kuder empleó un formato de reactivos de elección forzada al diseñar sus inventarios de intereses. Para elaborar su primer inventario, el Registro de Preferencias Vocacionales de Kuder, Kuder administró una lista de enunciados sobre actividades a estudiantes universitarios y, a partir de sus respuestas, determinó qué reactivos se agrupaban. Los resultados condujeron a la elaboración de diez grupos de reactivos con bajas correlaciones a través de los grupos de reactivos, pero con altas correlaciones dentro de los grupos. Los diez grupos de reactivos fueron: Exteriores, Mecánicos, Computacionales, Científicos, Persuasivos, Artísticos, Literarios, Musicales, de Servicio social y de Trabajo de oficina. Después se formaron tríadas de reactivos, donde cada miembro de una tríada pertenecía a un distinto grupo o área de interés, y se aplicaron en un formato de elección forzada. De esta manera, los intereses de las diez áreas distintas se confrontaron entre sí.

Los reactivos de los inventarios de Kuder consisten en tres enunciados de actividades, y se pide a los examinados responder qué actividad les gusta *más* (M) y cuál les gusta *menos* (m). Las respuestas del primer ejemplo de tríada de reactivos de la figura 12.1 (página 270) indican que esta persona preferiría visitar un museo y le gustaría menos ir a una galería de arte. De las actividades incluidas en la segunda tríada de reactivos, a la persona le agradaría más coleccionar monedas y menos coleccionar piedras.

El formato de reactivos de elección forzada tiene tanto ventajas como desventajas; aunque tiende a minimizar ciertos grupos de respuesta (aceptación, conveniencia social, etc.) a las personas en ocasiones les parece extraño. Otro problema es el carácter ipsativo de las respuestas a reactivos de elección forzada: al aceptar o rechazar una actividad en un área, quien responde el formato no selecciona ni rechaza una actividad de otra área. Por esta razón es imposible obtener calificaciones uniformemente altas o uniformemente bajas en todas las áreas de interés. Un patrón típico de calificaciones consiste en calificaciones altas en una o más áreas, calificaciones bajas en una o más áreas y calificaciones promedio en las áreas restantes.

En la actualidad, tres diferentes inventarios de Kuder están comercialmente disponibles: el Estudio de Intereses Generales de Kuder®, el Estudio de Intereses Ocupacionales de Kuder®, y la Búsqueda de Carrera de Kuder® (todos de National Career Assessment Services).

Estudio de Intereses Generales de Kuder®

Este inventario, que fue diseñado para los grados 6° a 12° y toma de 45 a 60 minutos en completarse, consiste en 168 tríadas de enunciados que describen diversas actividades; una actividad de

cada tríada debe marcarse como "Más preferida" y otra como "Menos preferida". Las respuestas se califican en diez áreas de interés general: Exterior, Mecánica, Computacional, Científica, Persuasiva, Artística, Literaria, Musical, Servicio social y Trabajo de oficina, además una escala de Verificación (V) indica si las respuestas se marcaron de manera cuidadosa. En 1987 se obtuvieron normas de rangos percentilares separadas para cuatro grupos (hombres y mujeres de los grados 6° a 8° y hombres y mujeres de los grados 9° a 12°). La disponibilidad de normas separadas por género permite a los examinados comparar sus calificaciones con las de niños y niñas. El formato de informe narrativo proporciona una lista de tablas de rangos percentilares por orden de clasificación en las diez áreas de interés, así como los tres temas del sistema RIASEC de Holland en que los examinados tienen mayor clasificación con respecto a otros hombres y a otras mujeres. Aunque la disposición de normas separadas por género ayuda a controlar el sesgo originado por normas combinadas, las diferencias en los intereses de hombres y mujeres pueden observarse en el hecho de que, en promedio, los niños tienen calificaciones más altas en las escalas Mecánica, Computacional, Científica y Persuasiva, mientras que las calificaciones de las niñas son más elevadas en las escalas Artística, Literaria, Musical, de Servicio social y de Trabajo de oficina.

Estudio de Intereses Ocupacionales de Kuder®

Este inventario, que fue diseñado para estudiantes de los grados 11° y 12°, estudiantes universitarios y adultos, consiste en 100 tríadas de enunciados que describen diversas actividades; una actividad de cada tríada debe marcarse como "Más preferida" y otra como "Menos preferida". Además de la forma tradicional de lápiz y papel, que toma entre 30 y 40 minutos en completarse, está disponible una versión del inventario en computadora. La calificación consiste en comparar las respuestas del examinando con las de personas que afirmaron estar satisfechas con sus elecciones ocupacionales; las respuestas también se comparan con las de estudiantes universitarios que están especializándose en campos de estudio particulares. La calificación de una persona en cualquiera de las escalas ocupacionales o especialidades universitarias es un coeficiente de correlación biserial modificado (*coeficiente lambda*) entre las respuestas dadas tanto por hombres como por mujeres a los reactivos y la proporción de individuos en el grupo específico ocupacional o de especialidad universitaria que apoyaron cada reactivo. Mientras más elevado sea el coeficiente lambda, más se asemejará la calificación de la persona al patrón de interés del grupo ocupacional o de especialidad correspondiente. Un informe narrativo de las calificaciones incluye los coeficientes lambda para escalas ocupacionales y de especialidad universitaria por género del grupo de norma. Los coeficientes lambda más elevados se destacan en la interpretación de la calificación, y las ocupaciones o especialidades asociadas son aquellas en que los intereses del examinando son mayores. Sin embargo, se recomienda que las ocupaciones o especialidades con coeficientes lambda dentro de las .06 unidades de los coeficientes mayores también se consideren.

Confiabilidad y validez

Los coeficientes de confiabilidad test-retest para corto plazo en los dos inventarios Kuder descritos líneas arriba están en el .80 y .90, y se ha encontrado que las calificaciones son bastante estables al cabo de una década o más (Zytowski, 1976). En general, la evidencia indica que la validez de contenido de ambos inventarios es satisfactoria. Con respecto a la validez predictiva del segundo inventario, Zytowski (1976) descubrió que más de la mitad de los individuos que lo habían resuelto de 12 a 19 años antes se habían incorporado a ocupaciones en las cuales habían tenido calificaciones de entre .07 y .12 en sus coeficientes lambda más elevados.

Búsqueda de Carrera de Kuder®

Este último miembro de la familia Kuder de inventarios de interés consta de 60 tríadas de reactivos de elección forzada escritos en un nivel de lectura de sexto grado. Fue diseñado para utilizarse desde el séptimo grado hasta la edad adulta y está disponible en inglés o español y en formatos basados en Internet, de retorno por correo y de autocalificación. Las calificaciones están basadas en las diez escalas de Kuder y en seis grupos de carreras.

La Búsqueda de Carrera de Kuder con Correspondencia para la Persona® incluye todas las características de la Búsqueda de Carrera de Kuder®, además de datos basados en un procedimiento de Correspondencia para la Persona®. Este procedimiento consiste en hacer corresponder las respuestas del examinando a las 60 tríadas de reactivos con las de "un fondo de criterios de personas reales con intereses y formas de hacer sus trabajos únicos". El informe de calificaciones incluye las 25 Correspondencias para la Persona® más cercanas a las respuestas del examinando. Los individuos que constituyen las 25 correspondencias superiores representan diversas carreras, pero tienen en común el hecho de que sus patrones de intereses son muy similares a los del examinando.

INTERESES Y PERSONALIDAD

De acuerdo con una concepción holística de la personalidad, los intereses y las habilidades son características de la personalidad. Por consiguiente, los inventarios de intereses también lo son de personalidad y los inventarios de personalidad también reflejan intereses (Holland, 1999). Más que desarrollarse por casualidad, los intereses son, de acuerdo con Darley y Hagenah (1955), reflejo o expresión de necesidades individuales y rasgos de personalidad profundamente enraizados. En concordancia con este punto de vista, una observación señala que, como sostienen los psicoanalistas desde la época de Freud, la selección vocacional está influida por rasgos de personalidad.

Teoría psicoanalítica sobre los intereses

Freud y otros psicoanalistas pusieron énfasis en los papeles de la *sublimación* (la canalización de impulsos sexuales o agresivos frustrados hacia actividades sustitutas) y la *identificación* (modelar el propio comportamiento con respecto a otra persona) en la formación de los intereses vocacionales. Con respecto a la sublimación, una persona con fuertes impulsos sádicos podría interesarse por convertirse ya sea en cirujano o en carnicero, una persona con intensas necesidades exhibicionistas podría convertirse en un actor o algún otro artista del escenario, y alguien cuyos impulsos sexuales están frustrados podría escribir poesía romántica o elegir una carrera relacionada con la decoración, la exhibición y otras áreas que ponen énfasis en el empleo del cuerpo (modelaje, actuación, deportes y actividades similares). Aunque la evidencia para la operación de la sublimación en determinar los intereses vocacionales y las elecciones profesionales está lejos de ser clara, los datos pertenecientes al papel de la identificación con los padres y otras personas significativas en la vida de una persona son más impresionantes (Crites, 1969; Heilbrun, 1969; Nachmann, 1960; Steimel y Suziedelis, 1963; Stewart, 1959).

La investigación que vincula los intereses vocacionales y educativos a necesidades y calificaciones específicas sobre otras características de la personalidad (por ejemplo, Utz y Korben, 1976) está relacionada, pero no queda restringida a la perspectiva psicoanalítica sobre intereses. Por ejemplo, se ha informado que las personas con intereses científicos tienden a ser más introvertidas, que el interés en vender está relacionado con la agresividad, y que las personas con in-

tensos intereses literarios y estéticos tienen más probabilidades de poseer características psiconeuróticas (Darley y Hagenah, 1955; Osipow, 1983; Super y Bohn, 1970). Otros investigadores (por ejemplo, Siegelman y Peck, 1960; Sternberg, 1955) han descrito patrones específicos de características de personalidad que están relacionados con la elección de una especialidad universitaria o carrera profesional.

Teoría de Roe e inventarios relacionados

Basada en cierta medida en teoría psicoanalítica y en la jerarquía de necesidades de Abraham Maslow (1954), así como en su propia investigación, Anne Roe (1956; Roe y Klos, 1969; Roe y Siegelman, 1964) concluyó que el principal factor en la elección de carrera es si el individuo está o no orientado hacia las personas. La teoría revisada de Roe contiene dos dimensiones independientes o continuos. En la primera dimensión ("orientación"), los papeles ocupacionales se clasifican en un rango que va desde la orientación hacia la comunicación con propósito, en un extremo, hasta la orientación hacia la utilización de recursos, en el otro extremo. En la segunda dimensión ("personas contra cosas"), los papeles ocupacionales van desde las relaciones interpersonales, en un extremo, hasta la orientación hacia los fenómenos naturales, en el otro extremo. Una tercera dimensión ("nivel"), bajo contra alto, consiste en el nivel de habilidad requerido por una ocupación (no capacitado, capacitado, profesional). Aunque estas tres dimensiones básicas son un rasgo central de la teoría de Roe, la teoría es en realidad mucho más elaborada y ha influido en el desarrollo de varios inventarios de intereses. Tres de estos instrumentos son el Inventario de Intereses COPS, el Inventario de Orientación Ocupacional de Hall y el Inventario de Intereses Vocacionales.

Inventario de Intereses COPS (COPS). Los ocho principales grupos de intereses que mide este inventario están representados alrededor del círculo que muestra la figura 12.3. El eje horizontal de la figura corresponde a la dimensión de las personas contra las cosas, el eje vertical corresponde a la dimensión de orientación, y la distancia desde el centro del círculo corresponde a la dimensión de nivel en el modelo de persona-ambiente de Roe. El COPS, cuyo tiempo de trabajo es de 20 a 30 minutos y requiere otros 15 a 20 minutos para la autocalificación, puede aplicarse a estudiantes a partir de la escuela secundaria.

Inventario de Orientación Ocupacional de Hall (HOOI). Este inventario, que se centra en 22 trabajos y características de personalidad, es adecuado para individuos desde el tercer grado hasta la edad adulta. Consiste en 112 reactivos de elección forzada que se centran en ocho áreas ocupacionales: Servicio, Contacto de negocios, Organización, Técnica, Exteriores, Ciencia, Cultura general, y Artes y entretenimiento.

Instrumentos y Teoría de Holland

Como se muestra en la figura 12.4 y fue descrito en la tabla 12.1, el modelo RIASEC de Holland (1985) conceptualiza las relaciones entre la personalidad y los intereses en términos de seis tipos de personalidades vocacionales. En correspondencia con estos seis tipos de personalidad hay seis ambientes modelo. Cada ambiente es buscado por personas con las habilidades, capacidades, actitudes, valores y rasgos de personalidad correspondientes.

De acuerdo con Holland, el comportamiento de una persona en un ambiente particular está determinado por la interacción entre personalidad y el tipo de ambiente. Las personas tienden a buscar ambientes congruentes con su personalidad y, en general, están más felices, más satis-

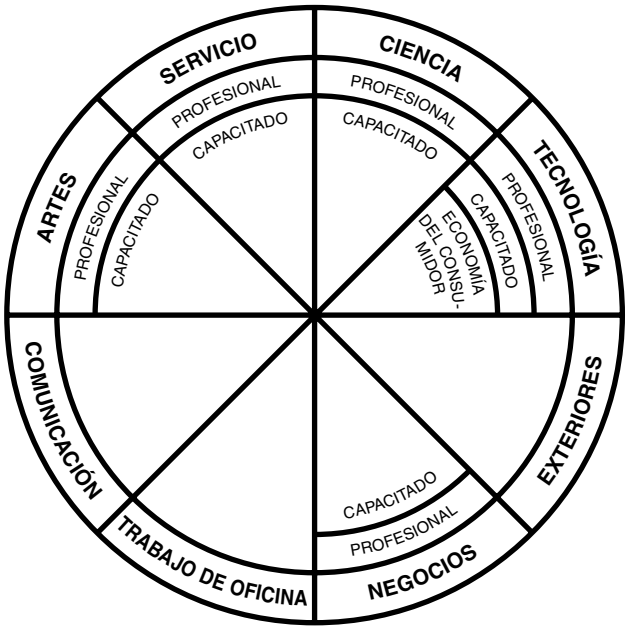


FIGURA 12.3 Ocho grupos de interés ocupacional evaluados por el Inventario de Intereses COPS.

(Reproducido con autorización de EdITS, San Diego, California 92107.)

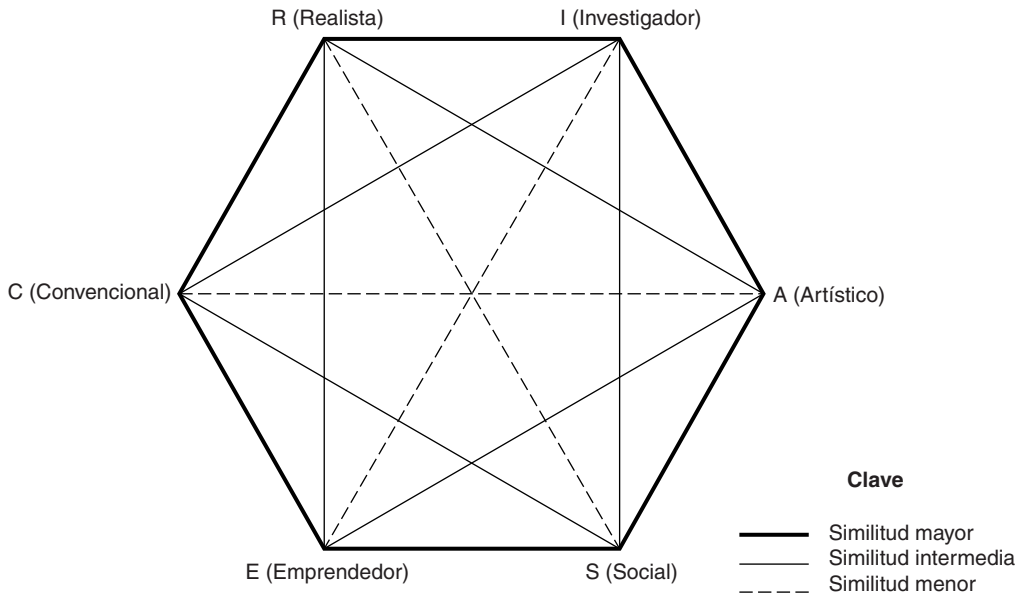


FIGURA 12.4 Modelo hexagonal de intereses, de Holland

fechas y son más productivas en esos ambientes que en otros que son incongruentes con su personalidad. Sin embargo, tanto los tipos de personalidad como los tipos de ambiente son idealizaciones y un individuo o ambiente dados regularmente están compuestos por una mezcla de más de un tipo ideal.

Algunos de los 15 pares de tipos de personalidad del modelo de Holland están más cercanamente relacionados entre sí y, por lo tanto, son más consistentes. La *consistencia* del patrón de interés de una persona está indicada por la medida en que obtenga calificaciones altas en los tipos de intereses que están cercanos entre sí en el modelo hexagonal. Debido a que los tipos investigador (I) y convencional (C) están más cerca del tipo realista (R) en el modelo, son más consistentes con éste. Por otra parte, los tipos artístico (A) y emprendedor (E) están más cercanos a, y por lo tanto son más consistentes con, el tipo social (S).

Otros conceptos importantes en la teoría de Holland comprenden diferenciación, identidad, congruencia y cálculo. Una persona con sólo una o dos calificaciones altas tiene un grado de *diferenciación* mayor que alguien con varias calificaciones elevadas. La elevada consistencia y diferenciación de los tipos de personalidad son característicos de las personas que manejan de manera eficiente sus problemas vocacionales. El segundo concepto, *identidad*, puede ser ya sea ambiental o personal. Un individuo con un sentido de *identidad personal* tiene una imagen clara y estable de sus metas, intereses y talentos; esta situación está relacionada con tener un reducido número de objetivos vocacionales en algunas categorías principales. La *identidad ambiental* se refiere a si las metas, tareas y recompensas del ambiente son estables a lo largo del tiempo. El tercer concepto, *congruencia*, se refiere al hecho de que distintos tipos de personalidad funcionan mejor en ambientes diferentes: ¿proporciona el ambiente oportunidades y recompensas congruentes con las habilidades y preferencias de la persona? Por último, el concepto de *cálculo* se relaciona con el hecho de que los tipos o ambientes de personalidad pueden ordenarse según un modelo hexagonal donde las distancias entre tipos o ambientes de personalidad sean consistentes con las relaciones teóricas entre ellos.

Durante muchos años, el modelo de personalidades vocacionales de Holland ha servido como estímulo y guía para la investigación sobre intereses y elecciones de carrera y, cuando se complementa con medidas de aspiración vocacional, ha alcanzado éxitos notables. Asimismo, las variables del modelo RIASEC han demostrado estar muy relacionadas con varias de las Cinco Grandes variables de la personalidad (Tokar y Fischer,1998); Tokar y Swanson, 1995 (vea el capítulo 17). Sin embargo, Holland (1996) ha sostenido que el modelo debería modificarse para incluir la idea de que distintos sistemas de creencias son característicos de tipos de personalidad diferentes y son promovidos por ambientes diferentes. Propuso aumentar el poder explicativo del modelo incorporándole de manera selectiva los conceptos de creencias y estrategias de carrera.

Investigación autodirigida. Hay dos inventarios de intereses vocacionales construidos con base en la teoría de Holland: Búsqueda Autodirigida e Inventario de Preferencias Vocacionales (de PAR). La Búsqueda Autodirigida (SDS) Forma R, que es uno de los inventarios de intereses de carrera más utilizado, consiste en un folleto de evaluación diseñado para ayudar al usuario a realizar una evaluación reflexionada sobre sus propios intereses y habilidades y un Buscador de Ocupaciones que ayuda a explorar de manera activa toda la gama de posibles ocupaciones. La SDS produce calificaciones en las seis variables RIASEC incluidas en la tabla 12.1 (página 274). Las normas se incorporan en un código ocupacional de tres letras, y el Buscador de Ocupaciones contiene más de mil títulos de ocupaciones vinculados al código. La SDS y el Buscador de Ocupaciones se han utilizado ampliamente en muchos contextos, incluyendo el Programa de Exploración de Carreras ASVAB del Departamento de Defensa de Estados Unidos (1999).

Inventario de Preferencias Vocacionales. El Inventario de Preferencias Vocacionales (VPI), un complemento de la Búsqueda Autodirigida y de otros inventarios de intereses, también está basado en la teoría de Holland acerca de que las ocupaciones pueden describirse en términos de características de personalidad. Quienes responden los inventarios indican si les agradan o desagradan cada una de las más de 160 diferentes ocupaciones del VPI, y sus respuestas se califican en las seis variables RIASEC y cinco escalas adicionales: Autocontrol, Estatus, Masculinidad-Feminidad, Infrecuencia y Aceptación. Las primeras seis calificaciones (RIASEC) pueden usarse con el Buscador de Ocupaciones SDS para propósitos de exploración de carreras y guía vocacional. Los distintos patrones de calificaciones altas en estos seis tipos originan la asignación del examinando a categorías vocacionales diferentes. Por ejemplo, una persona con calificaciones altas en las escalas Convencional, Empresarial y Social cae en la misma categoría como agente de publicidad y representante de ventas.

Mapa del Mundo Laboral

También basado en el modelo RIASEC de Holland, en combinación con las dimensiones de trabajo-tarea de Dale Prediger, el Mapa del Mundo Laboral se ilustra en la figura 12.5. El patrón de respuestas del Inventario de Intereses UNIACT del Programa Estadounidense de Evaluación Universitaria puede trazarse en este mapa y servir como base para ubicar y explorar regiones del mundo laboral. Grupos de trabajo similares (familias de trabajo) que cubren casi todo el panorama laboral de Estados Unidos están trazados en las 12 regiones del mapa. La ubicación de una familia de trabajo en el mapa se realiza con base en cuatro tareas de trabajo primarias:

Datos: Hechos, números, archivos, cuentas, procedimientos de negocios.
Ideas: Conocimiento, percepciones, teorías, nuevas formas de decir o hacer algo.
Personas: Servicios de cuidado, guía, ventas.
Cosas: Máquinas, herramientas, seres vivientes, y materiales tales como alimento, madera o metal.

Hacia el borde del mapa se anotan seis áreas generales del mundo laboral y los tipos RIASEC relacionados.

Diferencias de género

Los patrones de estereotipos sexuales de calificaciones obtenidas en el Inventario de Intereses de Strong, el Inventario de Preferencias Vocacionales, el Inventario de Intereses Vocacionales y otras medidas del interés son también reflejo de las diferencias de personalidad. Por ejemplo, las mujeres tienden a obtener calificaciones más altas que los hombres en los temas Social, Artístico y Convencional, y los hombres califican más alto que las mujeres en los temas Realista, Investigador y Emprendedor del VPI (Gottfredson, Holland y Gottfredson, 1975; Prediger y Hanson, 1976). Se han encontrado diferencias de género similares en las calificaciones del Inventario de Intereses Vocacionales y el Inventario de Intereses de Strong. Tales diferencias, como se argumenta, se deben a los reactivos que se centran en actividades o materiales específicos en los que un sexo tiene más experiencia que el otro. Ejemplos de ello son los reactivos de carpintería y reparación de automóviles de la escala Realista de Holland, que son más familiares para los hombres; rediseñar los reactivos para que incluyan una máquina de coser o una batidora podría volverlos más familiares para las mujeres.

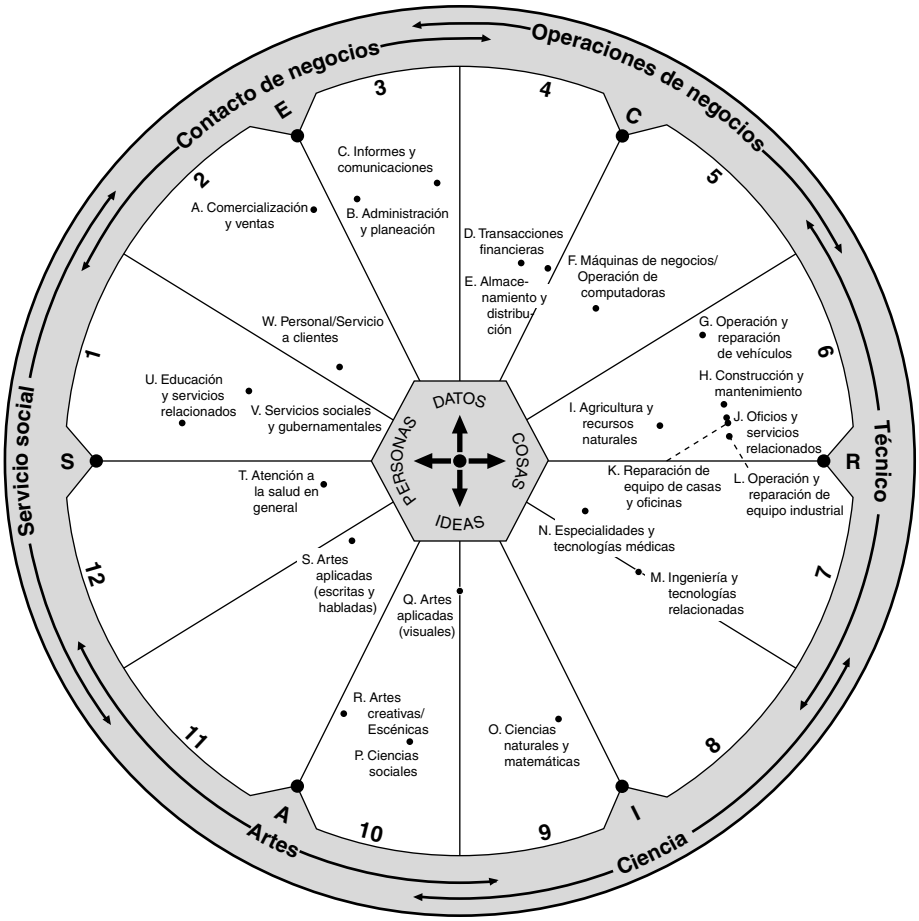


FIGURA 12.5 Mapa del Mundo Laboral, de ACT. Vea la explicación en el texto.

(Derechos reservados por ACT, Inc. Reproducido con autorización.)

Se ha afirmado, con cierta justificación, que las ediciones anteriores del inventario de Strong y de otros inventarios previos contribuyeron a la discriminación de género al dirigir a las jóvenes hacia ocupaciones tradicionalmente de mujeres, tales como la enseñanza básica, la enfermería y el trabajo de oficina (Diamond, 1979). Para responder a las acusaciones de sesgos por género, quienes elaboraron el Inventario de Intereses de Strong y algunos otros instrumentos psicométricos formularon versiones unisex de éstos. También proporcionaron normas de género combinadas, así como las tradicionales normas separadas por sexo. A fin de eliminar las diferencias por género en las respuestas, algunos instrumentos, tales como la versión revisada del Inventario de Intereses Vocacionales y el Inventario de Intereses UNIACT, emplean reactivos con equilibrio de sexos. Estos reactivos fueron los que eligieron aproximadamente porcentajes

iguales de hombres y mujeres. Tal vez una forma todavía más eficaz de reducir o eliminar las diferencias de género en las calificaciones de los inventarios de interés sea proporcionar oportunidades, estímulos y experiencias iguales para ambos sexos en diversas actividades, tanto en las que tradicionalmente son para hombres como en las que suelen ser para mujeres.[5]

Inventarios de personalidad e intereses

Los inventarios de personalidad como los descritos en el capítulo 17 contienen, en general, varios reactivos sobre intereses, actitudes y valores. Por esta razón, las calificaciones de muchos de estos inventarios proporcionan información útil para la consejería académica y vocacional. Uno de tales inventarios es el Cuestionario de 16 Factores de la Personalidad (16 FP). Los 16 FP pueden calificarse e interpretarse por computadora en variables ocupacionales que corresponden a los temas RIASEC de Holland. Se obtiene un perfil de calificaciones de estas variables que puede compararse con los perfiles de docenas de ocupaciones para determinar la similitud entre los intereses de quien responde el cuestionario y los de cada grupo ocupacional.

OTROS INVENTARIOS DE INTERESES CON PROPÓSITOS GENERALES Y ESPECIALES

Aunque los inventarios diseñados por Strong, Kuder y Holland han sido los más populares de todos los instrumentos psicométricos para evaluar intereses, se han elaborado muchas otras mediciones de intereses generales y con propósitos especiales. La mayoría de estos instrumentos se centran en intereses vocacionales, pero algunos se han diseñado principalmente para medir intereses en actividades relacionadas con la escuela y diversos tipos de pasatiempos. Un ejemplo es el Inventario de Búsqueda de Pasatiempos (por J. J. Liptak; JIST Works) que está diseñado para medir los intereses de pasatiempos de estudiantes a partir de secundaria. Los resultados pueden usarse como guía para elegir carrera o en asesoría laboral, y para ayudar a la persona a convertir sus intereses de pasatiempos en oportunidades de carrera o empleo.

En adición a los intereses estudiados, el Examen Campbell de Intereses y Habilidades (CISS) (de NCS Pearson) proporciona autoinformes de la confianza del examinando en su capacidad para desempeñar varias habilidades. También se han publicado muchos inventarios con propósitos especiales para evaluar los intereses de niños, gente discapacitada y personas que planean incorporarse a ocupaciones no profesionales.

Estudio de Intereses Vocacionales de Jackson

Uno de los inventarios de intereses generales que se han diseñado y validado con mayor cuidado es el Estudio de Intereses Vocacionales de Jackson (JVIS) (Sigma). Con base en los resultados de un amplio programa de investigación dirigido por D. N. Jackson, el JVIS consiste en 289 pares de enunciados de elección forzada que describen actividades relacionadas con el trabajo. Los enunciados que comprenden un reactivo con un par se refieren a dos intereses igualmente popu-

[5]La preocupación ante la discriminación por sexo, así como el carácter y el origen de las diferencias de sexo en las características psicológicas, también estimuló el desarrollo de numerosas mediciones del papel de los géneros, entre las cuales destaca el Inventario del Papel de los Sexos de Bem y el Cuestionario de Atributos Personales.

lares y se pide a los examinados indicar cuál interés prefieren. Diseñado para edades de bachillerato en adelante, el JVIS toma entre 45 y 60 minutos para completarse. La calificación inicial se basa en 34 escalas de interés básico que representan 26 dimensiones de papel de trabajo y 8 de estilo de trabajo. Las definiciones de estas dimensiones se perfeccionaron al referirse a las descripciones de trabajos del *Diccionario de títulos ocupacionales*. Otro método para calificar el JVIS se relaciona con 10 temas ocupacionales (Expresivo, Lógico, Inquisidor, Práctico, Asertivo, Socializado, Colaborador, Convencional, Emprendedor y Comunicación). Estos temas se basan en los seis temas de "personalidades vocacionales" de Holland y en un análisis factorial de las respuestas del JVIS. Las calificaciones de los 10 temas son bastante confiables: los coeficientes de consistencia interna varían de .70 a .92, y los coeficientes test-retest para un periodo de 4 a 6 semanas van de .69 a .92 para las escalas de interés básico. Con respecto a la validez del JVIS, un amplio estudio realizado en la Universidad Estatal de Pennsylvania descubrió que los perfiles del JVIS pronosticaban mejor la elección de especialidades académicas que cualquier otra combinación de mediciones de intereses y aptitudes anteriormente registrada.

Los revisores del JVIS han elogiado su cuidadosa elaboración y sus escalas factorialmente puras, pero se ha observado la necesidad de presentar más evidencias de su validez (Davidshofer, 1985; Thomas, 1985). Sin embargo, el manual revisado (Jackson, 2000) proporciona una cantidad considerable de documentación nueva concerniente a las características psicométricas del JVIS.

Inventarios para niños y personas con discapacidades

Debido a que los intereses vocacionales de los niños no suelen estar muy desarrollados ni ser realistas, se han diseñado inventarios de intereses sobre todo para individuos a partir de la escuela secundaria. Sin embargo, pueden aplicarse varios inventarios de intereses a niños de escuela elemental (grados 3° al 7°). Algunos ejemplos son el Inventario Intermedio de Orientación Ocupacional de HALL (de STS) y el Test de Intereses-Opiniones de Amplio Rango (WRIOT) (de Wide Range). Estos inventarios sirven para introducir y familiarizar a los niños con una amplia gama de actividades y ocupaciones en relación con sus intereses, experiencias, habilidades y aspiraciones presentes. El HALL Intermedio contiene 110 reactivos relacionados con la escuela, diseñados para complementar programas de desarrollo de conciencia y se centra en 22 características de trabajo y personalidad.

El WRIOT y algunos otros instrumentos (por ejemplo, el Inventario de Intereses Ilustrado de Geist y el Inventario de Intereses Vocacionales de Lectura, libre) se desarrollaron sobre todo para jóvenes con desventajas culturales y educativas. Más que palabras, frases, o enunciados, estos instrumentos emplean como materiales de prueba imágenes de personas que participan en ciertas actividades. El WRIOT consiste en 150 conjuntos de tres dibujos, presentados en un folleto o rollo de película, de personas que realizan diversas actividades y con las que pueden identificarse los individuos discapacitados física o mentalmente. El WRIOT puede terminarse en 40 a 60 minutos por individuos desde los cinco años de edad hasta adultos. Los gustos y aversiones se seleccionan en un formato de elección forzada y las respuestas se califican en grupos de 18 intereses (arte, ventas, administración, trabajo de oficina, equipos, operación de máquinas, atletismo, etc.) y ocho grupos de actitudes (sedentarismo, riesgo, ambición, estereotipo sexual y otros conceptos). Están disponibles listas de títulos de trabajos para cada uno de los 18 grupos de intereses.

En el inventario ilustrado de Geist (WPS) los examinados encierran en un círculo la imagen que prefieran de entre un grupo de tres ilustraciones. El inventario toma 25 minutos en completarse y se califica en 12 áreas. En el Inventario de Intereses Vocacionales de Lectura, libre (Elbern) los

examinados marcan preferencias para cada uno de los 55 conjuntos de tres dibujos que reproducen tareas de los trabajos. Las imágenes representan los tipos de tareas u ocupaciones donde los individuos con retraso mental pueden ser productivos y hábiles (automotrices, oficios de la construcción, trabajos de oficina, cuidado de animales, servicios de alimentos, cuidado de pacientes, horticultura, labores domésticas, atención personal, lavandería o manejo de materiales).

El inventario Evaluación de Intereses de Ashland (Sigma) fue diseñado para adolescentes y adultos con discapacidades de aprendizaje, problemas o retraso en el desarrollo, poca familiaridad con la lengua inglesa, acceso limitado a la educación, desempleo crónico, daño cerebral o trastornos emocionales o psiquiátricos crónicos. Los 144 pares de actividades relacionadas con las ocupaciones particulares que conforman este inventario se imprimen con letras grandes y en un grado de lectura de nivel tres. Los examinados seleccionan la actividad que más les gustaría realizar de las dos que incluye cada reactivo. Un ejemplo de tales reactivos es:

Limpiar tapetes	A
Entregar muebles	B

El inventario se califica en 12 escalas: Artes y oficios, Servicios de alimentos, Ventas, Servicios de protección, Mecánica, Cuidado de plantas o animales, Servicios personales, Trabajo de oficina, Servicios generales, Atención a la salud, Construcción y Transportes.

Intereses en ocupaciones no profesionales

Ciertas escalas del Inventario de Intereses de Strong, el Estudio de Intereses Ocupacionales de Kuder y otros inventarios de intereses generales corresponden a intereses en ocupaciones no profesionales, pero ninguno de estos inventarios se diseñó expresamente para tal fin. Desde la década de 1950 se han elaborado varios inventarios centrados en oficios especializados y son más sencillos en contenido y vocabulario que los inventarios de Strong y Kuder. Uno de tales instrumentos es el Inventario de Evaluación de Carreras (de NCS Pearson).

La Versión Vocacional del Inventario de Evaluación de Carreras (CAI) se realizó según el modelo de los inventarios de Strong y en ocasiones la han llamado "el Inventario de Intereses de Strong para el hombre trabajador". Las respuestas a los 305 reactivos del CAI se ubican en una escala de cinco puntos (A = le agrada mucho, a = le agrada un poco, I = le es indiferente, d = le desagrada un poco, D = le desagrada mucho). Los reactivos, escritos en un nivel de sexto grado y que cubren actividades, materias escolares y títulos de ocupaciones, pueden responderse en 30 a 45 minutos.

El informe basado en computadora de las calificaciones *T* del CAI consta de cuatro secciones: I. Índices administrativos (Respuestas totales, Consistencia de respuestas, Porcentajes en actividades, Materias escolares, Ocupaciones); II. Temas generales (Realista, Investigador, Artístico, Social, Emprendedor, Convencional); III. Escalas de áreas de intereses básicos, y IV. Escalas ocupacionales. Las 25 escalas de áreas de intereses básicos y las 111 escalas ocupacionales se agrupan en los seis temas de Holland. Se proporcionan calificaciones de las cuatro escalas especiales (Bellas Artes-Mecánica, Extroversión-Introversión ocupacional, Orientación educativa y Variabilidad de intereses), un informe narrativo generado por computadora y el resumen de un asesor. Desde un punto de vista psicométrico, el CAI está bien diseñado y tiene buena confiabilidad. También está disponible una versión ampliada del CAI, de 370 reactivos que puede calificarse en ocupaciones más profesionales que la versión vocacional. Tanto la versión vocacional como la ampliada pueden calificarse en temas ocupacionales generales con base en el modelo RIASEC de Holland.

UTILIZACIÓN DE LOS INVENTARIOS DE INTERESES EN LA CONSEJERÍA

A menudo hombres y mujeres jóvenes no son realistas al pensar en sus planes académicos y vocacionales. Muchos más estudiantes de bachillerato esperan graduarse de la universidad de los que realmente lo logran, y a menudo las aspiraciones profesionales de los graduados son incompatibles con sus posibilidades. Los objetivos y las decisiones con respecto a las carreras resultan afectados no sólo por factores ambientales, como la situación económica y el tipo de empleos disponibles, sino también por el género, la situación socioeconómica, las habilidades físicas y mentales, los intereses, y el conocimiento sobre ocupaciones en particular. Consecuentemente, los asesores de bachillerato y universidad deberían estar informados sobre el mundo laboral —lo que implican los trabajos particulares y si las habilidades, intereses y recursos que poseen los estudiantes son los adecuados para incorporarse o prepararse para determinados empleos. Los asesores vocacionales deben estar dispuestos a escuchar, informar y aconsejar; deben prepararse para obtener datos personales de los estudiantes y de quienes buscan su consejo, así como para proporcionarles información sobre programas de capacitación y ocupaciones.

La consejería vocacional debe llevarse a cabo con precaución. Tanto los empleos como los estudiantes son dinámicos y poseen múltiples facetas: tienen muchos rasgos distintos que además cambian con el tiempo. El trabajo de los psicólogos, por ejemplo, puede incluir diferentes actividades en distintas situaciones, y el carácter de dicho trabajo ha cambiado con los años. Incluso en el mismo contexto de situación y periodo, usualmente la mayoría de los trabajos tienen la suficiente diversidad como para que personas con distintas habilidades e intereses puedan adaptarse y desempeñarse de manera satisfactoria.

Intereses no son habilidades

Aunque persiste la necesidad de un enfoque flexible, probabilístico de la consejería vocacional, los asesores deben realizar cierta interpretación de las calificaciones de pruebas e inventarios psicológicos. Una distinción obvia se establece entre las aspiraciones y los intereses, por un lado, y las habilidades, por el otro. Aunque las calificaciones de inventarios de intereses vocacionales pueden tener correlaciones significativas con las mediciones de habilidades cognoscitivas, las correlaciones son, por lo general, bastante bajas (Hansen, 1984; Randahl, 1991; Swanson, 1993).

Las aspiraciones individuales suelen estar adornadas por percepciones falsas sobre lo que conlleva una actividad u ocupación particular. Por ejemplo, alguien que aspira a convertirse en enfermera puede verse como la famosa Florence Nightingale, ayudando a una humanidad doliente y disfrutando de los elogios y el amor de las personas. Es posible que no esté consciente de que la enfermería implica caminar y estar de pie durante mucho tiempo, vaciar orinales y escuchar quejas interminables. Por consiguiente, la eficiencia en la enfermería puede depender más de tener pies resistentes, tolerancia a los olores fuertes y una paciencia ilimitada, que de sentir amor por la humanidad.

Al interpretar los resultados de un inventario de intereses, los asesores deberían estar conscientes de la diferencia entre intereses y habilidades e intentar esclarecerla. Por lo general, una persona concluye, a partir de los resultados de un inventario de intereses, que tiene las habilidades necesarias para lograr el éxito en determinadas ocupaciones, cuando de hecho lo único que ha sido demostrado es que sus intereses son similares a los de las personas dedicadas a tal ocupación. Debido a que las personas a menudo no distinguen entre intereses y habilidades, no es sensato dejar la tarea de interpretar un inventario autoaplicado a la persona asesorada.

El hecho de que las correlaciones entre medidas de intereses y habilidades tienden a ser bastante bajas indica que muchas personas no poseen las habilidades requeridas para alcanzar el

éxito en los trabajos que les interesan, o que pueden no estar interesados en los trabajos para los que tienen las habilidades suficientes. Por esta razón, las calificaciones de los inventarios de intereses deberían emplearse en la consejería vocacional a la luz de otros datos sobre el asesorado. La información adicional que puede facilitar la decisión vocacional puede obtenerse de los logros del individuo (grados, premios, actividades extracurriculares, servicio a la comunidad y aspectos similares), experiencias y nivel de motivación.

Una variable motivacional con la que las habilidades, los intereses y el desempeño están moderadamente relacionados es la *autoeficacia*. En relación con la noción de autoconcepto, la autoeficacia se define como el juicio del individuo con respecto a su habilidad para realizar una tarea determinada en cierta situación. Betz (1992-1994) recomendaba que los consejeros vocacionales intentaran aumentar la autoeficacia del asesorado con respecto a un curso de trabajo u ocupación particular impulsando y preparando experiencias de éxito en áreas específicas. Para ayudar a esclarecer la relación circular entre la autoeficacia, el éxito y el interés, Lent, Lopez y Bieschke (1991) afirman que "las experiencias de éxito vividas en un dominio de desempeño particular pueden promover la autoeficacia; considerarse exitoso puede incrementar el interés en tal campo de dominio, y dicho interés motiva una mayor exposición y la elección de actividades vocacionales y educativas correspondientes" (p. 429).

Dadas las dificultades de explicar a estudiantes, padres y otras personas las relaciones del interés, las habilidades, la motivación, los logros pasados y el estatus financiero con el éxito vocacional y académico, los consejeros vocacionales experimentados deberían tener a su disposición una gran cantidad de fuentes de información y explicación concernientes al mundo laboral. Además de dichos libros de referencia, como el *Dictionary of Occupational Titles* (*DOT*) (Departamento del Trabajo de Estados Unidos, 1991, 1993),[6] el *Occupational Outlook Handbook, 2000-2001* (Oficina de Estadísticas Laborales, 2000) y el *Occupational Outlook for Colleges Graduates* (Oficina de Estadísticas Laborales, 1996), los consejeros vocacionales deberían tener a su disposición diversos materiales de investigación y desarrollo de carreras publicados por compañías comerciales de evaluación y por el Departamento del Trabajo de Estados Unidos. Gran parte de la información sobre carreras que antes estaba disponible sólo en libros puede obtenerse actualmente en Internet de la Oficina de Estadísticas Laborales y otras organizaciones públicas y privadas. Ejemplos de tales recursos son el *Occupational Outlook Handbook* (stats.bls.gov/ocohome), la *Career Guide to Industries, edición 2000-01* (stats.bls.gov/cghome.htm) y la *Occupational Outlook Quarterly* (stats.bls.gov/opub/ooq/ooqhome.htm). Dos sitios útiles para estudiantes universitarios son el BLS Career Information (www.bls.gov/K12/html/edu_over.htm), el cual proporciona a los alumnos de escuela elemental una introducción a la información de guía de carreras, y el *Career Planning Survey* (www.act.org/cps/index.html) de ACT, un amplio programa de guía de carreras para preparar a estudiantes de los grados 8° al 10° y que se informen para tomar decisiones sobre sus carreras y educación en los inicios del bachillerato.

Consejería vocacional basada en la computadora

Hay disponibles varios sistemas de guía asistidos por computadora que han sido diseñados para ayudar a los estudiantes a explorar sus intereses, valores, actitudes y habilidad para tomar decisiones de carrera realistas. Como ejemplos de programas de guía de carreras se encuentran

[6]O*NET, la Red de Información Ocupacional, una base de datos que incluye información sobre conocimientos, habilidades, capacidades, intereses, preparación, contextos y tareas relacionadas con 1,122 ocupaciones, recientemente ha reemplazado al *Dictionary of Occupational Titles* (vea el sitio de Internet www.doleta.gov/programs/onet/).

SIGI PLUS de ETS (vea la página Web www.ets.org/sigi/index.html), DISCOVER de ACT (página Web www.act.org/discover), Estudio de Planeación de Carrera de ACT (página Web www.act.org/cps/indes.html), Sistema de Planeación de Carrera de los Servicios Nacionales de Evaluación de Carreras (página Web www.kuder.com/kcps.asp), y Sistema de Información de Carreras de la Universidad de Oregon (página Web oregoncis.uoregon.edu).

El Sistema de Información y Guía Interactivo (SIGI PLUS) es un programa interactivo de cómputo que fue desarrollado por el Servicio de Evaluación Educativa (ETS) para ayudar a los usuarios a tomar decisiones sobre su carrera en forma racional e informada. De acuerdo con el ETS, SIGI PLUS "ayuda a los usuarios a evaluar sus valores, intereses, habilidades y recursos, y a relacionar estas características con las recompensas, satisfacciones, actividades y requisitos que buscan en una carrera". Sin embargo, como el ETS considera que los intereses expresados son por lo menos tan precisos como los inventarios de intereses, SIGI PLUS no incluye un inventario de intereses. Similares a SIGI PLUS son la Guía de Carreras DISCOVER y el Sistema de Programación de Información (para bachillerato, universidades y organizaciones de adultos), y la Versión de Escuela Intermedia DISCOVER. Según se informa, los usuarios consideran a estos programas de cómputo divertidos y útiles, y después de varias sesiones es posible que se llegue al análisis con un consejero vocacional experimentado. Los estudiantes universitarios y los asesores en capacitación han asignado calificaciones altas a SIGI PLUS y DISCOVER (Kapes, Borman y Frazier, 1989; Peterson, Ryan-Jones, Sampson, Readon, *et al.*, 1994).

RESUMEN

Los inventarios son los más populares de todos los métodos existentes para evaluar intereses. El uso serio de inventarios de intereses en la consejería y colocación vocacional y académica empezó con la elaboración del Formulario de Intereses Vocacionales de Strong en las décadas de 1920 y 1930. La publicación subsecuente del Registro de Preferencias Vocacionales de Kuder y de otros inventarios de intereses dio origen a que aumentara la comprensión de los intereses y mejorara la precisión con que pueden medirse.

Las calificaciones de los inventarios de intereses no sirven para pronosticar muy bien el éxito vocacional, pero sí funcionan satisfactoriamente al pronosticar la elección y satisfacción ocupacional. Los resultados de estudios longitudinales han demostrado que los intereses inventariados son bastante estables, aunque las respuestas a dichos inventarios pueden ser fingidas y son susceptibles a provocar grupos de respuesta. En general, se considera que los intereses se aprenden, pero hay ciertas evidencias de una base hereditaria para las preferencias de diferentes tipos de personas, actividades y cosas.

La versión más reciente del Inventario de Intereses de Strong consta de 317 reactivos agrupados en ocho categorías. Se califica en seis temas ocupacionales generales, 25 escalas de interés básico, 211 escalas ocupacionales, cuatro escalas de estilo personal y tres índices administrativos. Se desarrollaron claves de calificación de las escalas ocupacionales de manera empírica al comparar las respuestas de las personas en general con las de personas empleadas en las ocupaciones particulares. Las calificaciones del Inventario de Intereses de Strong son formas bastante confiables y válidas de pronosticar la persistencia y satisfacción ocupacionales, pero no necesariamente el éxito ocupacional.

G. F. Kuder elaboró varios inventarios de intereses consistentes en una serie de tríadas de reactivos de elección forzada. El Registro de Preferencias Vocacionales de Kuder, el Estudio de Intereses Generales de Kuder y el Estudio de Intereses Ocupacionales de Kuder son tres instrumen-

tos de este tipo. El último de ellos, como el Inventario de Intereses de Strong, puede calificarse en varias escalas ocupacionales empíricamente derivadas, pero los primeros dos se calificaron en escalas de intereses generales. Todos estos inventarios, de nuevo igual que el Inventario de Intereses de Strong, son adecuados para estudiantes de bachillerato y adultos. Recientemente se ha sumado a la familia de instrumentos de Kuder, la Búsqueda de Carreras de Kuder.

Las relaciones entre intereses y personalidad se destacan en la investigación de Anne Roe y John Holland y los instrumentos derivados de su investigación. Búsqueda Autodirigida e Inventario de Preferencias Vocacionales, ambos de Holland, son dos de los inventarios de intereses más populares. El modelo de intereses RIASEC, de Holland, también ha influido en el desarrollo y la calificación de algunos otros inventarios de intereses.

Además de los inventarios calificados de acuerdo con amplias áreas de interés y de ocupaciones de adultos, están disponibles algunos instrumentos para evaluar los intereses de los niños, de las personas con desventajas culturales o discapacidad, y de quienes planean incorporarse a ocupaciones no profesionales.

La consejería vocacional requiere de un amplio conocimiento del mundo laboral y de una integración experimentada de calificación de pruebas de habilidad, mediciones de intereses, datos biográficos y observaciones de la conducta.

PREGUNTAS Y ACTIVIDADES

1. ¿Qué son los grupos de respuesta? ¿Por qué preocupan particularmente al diseñar inventarios de intereses y características de personalidad? ¿Qué puede hacerse para contrarrestar los efectos de los grupos de respuesta en las calificaciones de estos inventarios?

2. En colaboración con su instructor responda el Inventario de Intereses de Strong, el Estudio de Intereses Ocupacionales de Kuder o la Búsqueda Autodirigida de Holland y haga que lo califiquen. Después de recibir su informe de calificaciones, busque la colaboración de su instructor o asesor para interpretar los resultados.

3. Compare el Inventario de Intereses de Strong (SII) con el Estudio de Intereses Ocupacionales de Kuder (KOIS) en términos de diseño, calificación e interpretación. ¿Por qué a menudo las correlaciones entre las escalas ocupacionales del SII y el KOIS que tienen los mismos nombres o nombres muy similares no son muy altas? ¿Qué significado teórico y práctico podría tener esto?

4. Considere los inventarios de interés:
- **a.** Inventario de Evaluación de Carreras
- **b.** Estudio de Intereses Ocupacionales de Kuder
- **c.** Búsqueda Autodirigida
- **d.** Inventario de Intereses de Strong
- **e.** Prueba de Rango Amplio de Interés-Opinión

¿Cuál recomendaría para aplicar en las siguientes situaciones?

(1) Asesoría de un estudiante universitario del primero o último años sobre la elección de su especialidad y vocación.
(2) Establecer un programa de asesoría para estudiantes que ingresan a un bachillerato vocacional con diversos programas de oficios.
(3) Ayudar a los alumnos a investigar carreras y considerar diversas opciones de ocupación.
(4) Ayudar a un grupo de individuos discapacitados física y mentalmente a considerar diversas actividades con las que podrían identificarse.

(5) Ayudar a un servicio de asesoría a trabajar con graduados universitarios que no están seguros sobre sus carreras futuras.
(6) Introducir y familiarizar a un grupo de niños de escuela elemental con un amplio rango de actividades de carrera y ocupaciones.

5. Defienda la tesis de que los intereses son características de la personalidad y que, por lo tanto, los inventarios de intereses son mediciones de la personalidad. Cite teorías específicas, hallazgos de investigaciones e instrumentos para apoyar su posición.

6. Escriba la letra del área de interés de la columna de la derecha que corresponda a la característica de personalidad de la columna izquierda. Puede elegir más de una letra de la columna derecha frente al adjetivo de la columna izquierda.

1. activo	**a.** artístico
2. agresivo	**b.** trabajo de oficina
3. atractivo	**c.** computacional
4. encantador	**d.** literario
5. conservador	**e.** mecánico
6. minucioso	**f.** musical
7. convencional	**g.** exteriores
8. cooperativo	**h.** persuasivo
9. valiente	**i.** científico
10. dependiente	**j.** servicio social
11. amistoso	
12. obstinado	
13. inteligente	
14. inventivo	
15. lógico	
16. metódico	
17. organizado	
18. original	
19. extrovertido	
20. agradable	
21. práctico	
22. preciso	
23. robusto	
24. serio	
25. sociable	
26. fuerte	
27. compasivo	
28. táctico	
29. reflexivo	
30. ahorrativo	

Compare sus respuestas con las de sus compañeros de clase y amigos. ¿Hubo alguna consistencia en la forma en que distintas personas relacionaron los reactivos de la columna izquierda con los de la derecha? ¿Qué le indican los resultados sobre las relaciones entre la personalidad y los intereses vocacionales? ¿Algunas características de la personalidad están relacionadas con los intereses en una amplia variedad de ocupaciones, mientras que otras se asocian con una gama más estrecha de ocupaciones?

7. En la teoría sobre intereses y personalidad de J. L. Holland, ciertos grupos de intereses son característicos de personas con determinados rasgos de personalidad e indican determinadas vocaciones o carreras. Complete las partes I y II respondiendo los seis enunciados incluidos y luego determine en la Parte III con cuánta precisión sus respuestas pronosticaron las ocupaciones en que cree usted estar verdaderamente interesado.

Parte I. Intereses

¿Cuál de las siguientes descripciones se ajusta más a usted: A, B, C, D, E o F?

A. Le gusta y es bueno para manipular herramientas, máquinas y otros objetos y trabajar en exteriores con plantas y animales.

B. Le gusta y es bueno para observar, aprender, investigar, analizar, evaluar y resolver problemas.

C. Le gusta y se desempeña bien en situaciones no estructuradas en que su creatividad o imaginación pueden expresarse.

D. Le gusta y es bueno para trabajar con otras personas; desarrollar, inspirar, informar, capacitar, curar, ayudar o apoyar de diversas maneras a los demás.

E. Le gusta y es bueno para influir y convencer a otras personas, y le gustaría dirigir o administrar una organización.

F. Le gusta y es bueno para actividades aritméticas y de trabajo de oficina, tal como archivar, llevar registros y procesar datos.

Parte II. Características de personalidad

¿Cuál de las siguientes descripciones concuerda más con usted: A, B, C, D, E o F?

A. Usted es una persona realista, práctica, adaptable y natural.

B. Usted es una persona racional, cautelosa, curiosa, independiente e introvertida.

C. Usted es una persona imaginativa, introspectiva, complicada, emocional, expresiva, impulsiva, no conformista y desordenada.

D. Usted es una persona cooperativa, amistosa, colaboradora, convincente, táctica y comprensiva.

E. Usted es una persona emprendedora, ambiciosa, energética, dominante, hedonista, segura de sí misma, sociable y platicadora.

F. Usted es una persona meticulosa, eficiente, inflexible, obediente, ordenada, persistente y autocontrolada.

Parte III. Vocaciones sugeridas

Los siguientes enunciados corresponden a la(s) letra(s) que seleccionó en las partes I y II y son trabajos que con más probabilidad podrían agradarle.

A. Trabajos como mecánico automotriz, granjero o electricista.

B. Trabajos en campos como química, física, biología, geología y otras ciencias.

C. Trabajos como actor, músico o escritor.

D. Trabajos en áreas de asesoría psicológica, terapia de lenguaje, enseñanza y otras afines.

E. Trabajos como administrador, ejecutivo de ventas o vendedor.

F. Trabajos como banquero, bibliotecario y experto en impuestos.

¿Sus respuestas a las partes I y II indicaron las mismas vocaciones en la parte III? ¿Estas vocaciones sugeridas son consistentes con las que siente que verdaderamente le interesan y para las que tiene las capacidades y personalidad necesarias con las cuales alcanzar el éxito?

8. ¿Qué efectos cree usted que puede tener la aplicación y calificación de inventarios de intereses vocacionales por medio de la computadora o Internet en los procesos de familiarización, búsqueda, asesoría y toma de decisiones relacionados con las carreras?

9. Conéctese a las siguientes tres páginas Web y explore la información que se proporciona en relación con las carreras:

www.bls.gov/k12/htm/edu_over.htm
www.act.org/cps/index.html
www.kuder.com/kcps.asp

CAPÍTULO TRECE

ACTITUDES, VALORES Y ORIENTACIONES PERSONALES

Una *actitud* es una predisposición aprendida para responder positiva o negativamente ante un objeto, una situación, institución o persona en particular. Como tal, consta de componentes cognoscitivos (de conocimiento o intelectuales), afectivos (emocional y motivacional) y de desempeño (conductual o de acción). Aunque el concepto de actitud es similar en ciertos aspectos al de interés, opinión, creencia o valor, hay diferencias en la forma en que se usan estos términos. Un *interés* es un sentimiento o preferencia relativo a nuestras propias actividades. A diferencia de una *actitud*, que implica aprobación o desaprobación (un juicio moral), estar interesado en algo simplemente implica que una persona dedica tiempo a pensar o a reaccionar ante ello, sin importar si estos pensamientos y comportamientos son positivos o negativos. Una *opinión* es una reacción específica ante ciertos sucesos o situaciones, mientras que una *actitud* es más general en cuanto a sus efectos en las respuestas a un amplio rango de personas o acontecimientos. Además, la gente es consciente de sus opiniones, pero probablemente no perciba del todo sus actitudes. Las opiniones son similares a las *creencias* en cuanto a que también son juicios o aceptación de ciertas proposiciones como hechos consumados, pero los soportes reales para sustentar las opiniones suelen ser más débiles que los utilizados para afirmar las creencias. Por último, el término *valor* se refiere a la importancia, utilidad o mérito asociado a determinadas actividades y objetos, por lo regular como fines, pero potencialmente como medios también.

MEDICIÓN DE ACTITUDES

Pueden usarse distintos métodos para obtener información concerniente a la actitud de una persona hacia algo, incluyendo la observación directa, las técnicas proyectivas, los indicadores fisiológicos, las mediciones de las asociaciones implícitas y los inventarios o escalas de actitud. Tal vez el procedimiento más inmediato sea la observación directa, es decir, observar cómo se comporta la persona en relación con ciertas cosas. ¿Qué hace la persona realmente en situaciones donde el objeto o acontecimiento de la actitud está presente? La disposición a hacer un favor, firmar una petición y realizar una donación a alguna causa son ejemplos de mediciones conductuales de actitudes.

Técnicas proyectivas, fisiológicas e implícitas

La observación directa del comportamiento es informativa, en particular con niños pequeños o cuando otros métodos se consideran una intrusión. No obstante, obtener una muestra represen-

tativa del comportamiento a través del tiempo y las situaciones puede requerir un periodo prolongado y resultar costoso. Asimismo, las mediciones conductuales de las actitudes a menudo producen resultados de técnicas proyectivas y cuestionarios o escalas de actitudes. Por varias razones, una persona puede desempeñar un papel o comportarse de otras maneras engañosas y sugerir así actitudes diferentes de las que en realidad posee.

Técnicas proyectivas. Las actitudes pueden evaluarse de manera proyectiva mostrando un conjunto de imágenes ambiguas a las personas e indicándoles que inventen una historia sobre cada imagen. Debido a que los dibujos pueden interpretarse de diversas formas, las historias que relatan las personas deben revelar algo sobre sus actitudes hacia los personajes, escenas o situaciones de las imágenes. Dos técnicas proyectivas que pueden usarse para determinar actitudes individuales son las asociaciones de palabras y el completar enunciados (vea el capítulo 18). La información indirecta sobre las actitudes de una persona también puede obtenerse en una prueba de conocimiento sobre los hechos correspondientes a dichas actitudes (Korman, 1974; Moyer, 1977).

Técnicas fisiológicas. Otro procedimiento que se ha empleado en ocasiones para medir las actitudes hacia diversos estímulos consiste en vigilar las reacciones, tales como los cambios en la conducción eléctrica de la piel del individuo (GSR o respuesta galvánica de la piel) o las modificaciones en el diámetro de la pupila del ojo en respuesta a estímulos relacionados con actitudes. Por ejemplo, el prejuicio contra ciertos grupos étnicos se ha medido por cambios en la magnitud de la GSR y del diámetro de la pupila en respuesta a fotografías de personas pertenecientes a dichos grupos (por ejemplo, Cooper y Pollock, 1959; Hess, 1965; también vea Wingard y Maltzman, 1980). Sin embargo, estas técnicas fisiológicas no son muy eficientes y, más que evaluar las actitudes en sí mismas, pueden estar midiendo excitación, interés o la respuesta de orientación (Woodmansee, 1970).

Asociaciones implícitas. Incluso cuando las actitudes, o bien otros pensamientos y creencias, no se revelan explícitamente en el comportamiento de una persona, pueden estar implícitas en lo que dice o hace. Métodos tales como la Prueba de Asociación Implícita de ventana de respuesta (Greenwald, McGhee y Schwartz, 1998), una tarea de reacción automática en el tiempo que mide el sesgo evaluador, puede revelar prejuicios implícitos u otras tendencias de respuesta negativa o positiva, aunque no se manifiesten explícitamente en los actos de la persona (por ejemplo, Rudman, Grenwald, Mellott y Schwartz, 1999). Dichas actitudes implícitas con frecuencia no pueden captarse mediante la introspección ni siquiera por la persona que reflexiona.

Escalas tradicionales de actitud

El método más popular de medición de actitudes es aplicar una *escala de actitudes* consistente en un conjunto de enunciados positivos y negativos concernientes a un concepto específico (un grupo de personas, una institución, un concepto). La calificación total de una escala de actitud se determina a partir de las respuestas agregadas de los examinados a los enunciados, con el método específico de calificación que depende del tipo de escala.

Una de las primeras escalas de actitud fue la Escala de Distancia Social de Bogardus (Bogardus, 1925), en la que los examinados clasificaban varios grupos raciales y religiosos en orden de aceptación. Al aplicar esta *escala acumulativa* se indicaba a las personas que señalaran

el grado en que aceptaban a varios grupos sociales o religiosos en diversas habilidades. Los reactivos se ordenaban en una jerarquía tal, que una respuesta positiva a un reactivo dado implicaba respuestas positivas para todos los reactivos anteriores en la jerarquía. La escala Bogardus probó ser útil en la investigación sobre diferencias regionales y otras variables relacionadas con el prejuicio racial, pero permitía la medición de actitudes sólo en una escala ordinal y era más bien cruda considerada según parámetros actuales. Se originaron mejores mediciones de actitud a partir de la investigación de Louis Thurstone, Rensis Likert, Louis Guttman y otros expertos en mediciones psicométricas.

Escalas tipo Thurstone. Hacia finales de la década de 1920, Louis Thurstone y sus colegas intentaron medir las actitudes con una escala de medición de intervalo, en la cual las diferencias iguales en los valores de la escala corresponden a iguales diferencias en la fuerza de la actitud, usando los métodos de *comparaciones de pares* e *intervalos de igual aparición*. La elaboración de una escala de actitud mediante cualquiera de estos métodos empieza por recopilar un gran número de enunciados que expresen un amplio rango de sentimientos positivos y negativos hacia un tema dado. El siguiente paso en el método de comparaciones de pares es que muchos expertos comparen los enunciados entre sí e indiquen cuál enunciado de cada par expresa una actitud más positiva hacia el tema. Debido a que realizar las numerosas comparaciones que requería este procedimiento es más bien engorroso y se lleva mucho tiempo, el método de intervalos de igual aparición resultó más popular.

En el método de intervalos de igual aparición, los aproximadamente 200 enunciados que expresan actitudes hacia algo (persona, objeto, evento, situación o abstracción) se dividen en 11 categorías por una muestra grande de jueces. Estas categorías van desde la menos favorable (categoría 1) hasta la más favorable (categoría 11) hacia el asunto de que se trate. A los jueces se les indica que consideren las 11 categorías como ubicadas a intervalos iguales a lo largo de un continuo. Una vez que todos los jueces hayan completado el proceso de clasificación de todos los enunciados, se elabora una distribución de frecuencia para cada enunciado contando el número de jueces que ubicaron el enunciado en cada categoría. A continuación se calculan la mediana (*valor de escala*) y el rango semiintercuartilar (*índice de ambigüedad*) a partir de la distribución de frecuencia correspondiente. Entonces, los enunciados se clasifican por orden de sus valores de escala, y se seleccionan alrededor de 20 enunciados para la escala terminada. En una escala de intervalo verdadera, la diferencia entre los valores de escala de cualquier par de enunciados adyacentes será igual a la diferencia entre los valores de escala de cualquier otro par de enunciados adyacentes. Además, los índices de ambigüedad de todos los enunciados deberán ser bajos.

En la forma 13.1 se muestra parte de una de las múltiples escalas de actitud construida mediante este método. Obsérvese que los valores de escala de los enunciados, que se refieren a actitudes hacia la pena de muerte, van desde .1 (muy negativa hacia) hasta 11.0 (muy positiva hacia). La calificación de una persona en una escala de ese tipo es la mediana de los valores de escala de los enunciados que elija.

Thurstone y sus colegas elaboraron aproximadamente 30 inventarios de actitud mediante el método de intervalos de igual aparición, la mayoría de los cuales tenía confiabilidad del orden de .80. Remmers (1960) generalizó el procedimiento de intervalos de igual aparición en sus nueve Escalas de Actitud Dominante, que miden las actitudes hacia cualesquier materia escolar, vocación, institución, grupo definido, acción social propuesta, práctica, actividad doméstica, conducta individual y de grupo, y el bachillerato.

FORMA 13.1 Doce de los veinticuatro reactivos de la escala de actitudes hacia la pena de muerte

Instrucciones: Éste es un estudio de actitud hacia la pena de muerte. A continuación se presentan varios enunciados que expresan distintas actitudes hacia la pena de muerte.

✓ Ponga una marca en esta forma si está de acuerdo con el enunciado.
× Ponga una cruz si está en desacuerdo con el enunciado.

Trate de indicar acuerdo o desacuerdo para cada enunciado. Si simplemente no puede decidir sobre un enunciado, puede escribir un signo de interrogación. Éste no es un examen. No hay respuestas correctas ni incorrectas para estos enunciados. Se trata solamente de un estudio sobre las actitudes de las personas hacia la pena de muerte. Por favor, indique sus propias convicciones mediante la marca correspondiente para el caso de estar de acuerdo o en desacuerdo.

Valor de escala	Número de reactivo	
(0.1)	12	No creo en la pena de muerte en ninguna circunstancia.
(0.9)	16	La ejecución de criminales es una vergüenza para la sociedad civilizada.
(2.0)	21	El Estado no puede enseñar lo sagrado de la vida humana destruyéndola.
(2.7)	8	La pena de muerte nunca ha sido eficaz para evitar el crimen.
(3.4)	9	No creo en la pena de muerte, pero no estoy seguro de que no sea necesaria.
(3.9)	11	Creo que el regreso del poste de flagelación sería más efectivo que la pena de muerte.
(5.8)	18	No creo en la pena de muerte, pero no es práctico abolirla.
(6.2)	6	La pena de muerte está mal, pero es necesaria en nuestra civilización imperfecta.
(7.9)	23	La pena de muerte se justifica sólo por asesinato premeditado.
(9.4)	20	La pena de muerte le da su merecido al criminal.
(9.6)	17	La pena de muerte es justa y necesaria.
(11.0)	7	Todo criminal debería ser ejecutado.

(Reimpreso de Peterson y Thurstone, 1933.)

A pesar de la confiabilidad bastante alta de los instrumentos elaborados mediante el método de intervalos de igual aparición, el procedimiento se ha criticado por todo el trabajo que implica la elaboración de la escala, la falta de originalidad de la calificación de una persona, y los efectos de las propias actitudes de los jueces sobre los valores de la escala o los enunciados (Selltiz, Wrightsman y Cook, 1976). Considerando la gran disponibilidad de las computadoras y otros dispositivos para ahorrar tiempo, la primera crítica no es tan seria. La segunda crítica se refiere al hecho de que la misma calificación, que es simplemente la mediana del valor de escala de los enunciados con los que se estuvo de acuerdo, puede obtenerse eligiendo una combinación de enunciados distinta y, por lo tanto, no es única. La tercera crítica es relativa al hecho de que no todas las personas son capaces de desempeñar el papel de juez imparcial. En un estudio (Goodstadt y Magid, 1977) se descubrió, por ejemplo, que casi 50% de los estudiantes universitarios que actuaron como jueces respondieron a un conjunto de enunciados de actitud en términos de su propio acuerdo o desacuerdo personal con los reactivos. Sin embargo, Bruvold (1975) concluyó que dar las instrucciones cuidadosamente a los individuos que se encargan de juzgar disminuye el sesgo del juicio a un nivel que no distorsiona gravemente las propiedades de intervalo

de igual aparición de estas escalas. Una última crítica, que no se limita a las escalas de actitud tipo Thurstone, es que las calificaciones de estas escalas representan mediciones sólo en un nivel ordinal y no intervalar. De hecho, el nivel de la medición de escalas de actitud elaboradas por el método de escalas de intervalo de igual aparición probablemente se encuentra en algún punto entre los niveles ordinal y de intervalo de medición.

Escalas tipo Likert. El más popular de todos los procedimientos de escalamiento de actitud, sin duda debido a su sencillez y versatilidad, es el procedimiento diseñado por Rensis Likert. Al igual que con el método Thurstone de intervalos de igual aparición, el *método de rangos sumarizados* empieza con la recopilación o elaboración de una gran cantidad de reactivos de enunciados que expresen diversas actitudes positivas y negativas hacia un objeto o acontecimiento específico. En la tabla 13.1 se presentan sugerencias para elaborar enunciados de actitud.

Después de diseñar un conjunto de enunciados preliminares, se indica a un grupo de 100 a 200 personas seleccionadas, no necesariamente jueces expertos, que en una escala de 4 a 7 puntos señalen la medida en que están de acuerdo o en desacuerdo con cada enunciado. En el caso típico de una escala de 5 puntos, los reactivos expresados en forma positiva se califican con 0 para muy en desacuerdo, 1 para en desacuerdo, 2 para indeciso, 3 para de acuerdo y 4 para muy de acuerdo; los reactivos expresados de manera negativa se califican con 4 para muy en desacuerdo, 3 para en desacuerdo, 2 para indeciso, 1 para de acuerdo y 0 para muy de acuerdo. La calificación total de la persona en el conjunto inicial de reactivos se calcula como la suma de sus calificaciones en los reactivos individuales. Después de obtener las calificaciones totales para todas las personas que respondieron en el conjunto de reactivos inicial, se aplica a cada reactivo un procedimiento estadístico (prueba *t* o índice de discriminación de los reactivos). Entonces se seleccionan cantidades iguales de reactivos expresados de manera positiva y negativa (por lo regular, diez de cada uno) que distingan de manera significativa a los participantes cuyas calificaciones totales correspondan al 27% superior de aquellos cuyas calificaciones se ubiquen en el 27% inferior. En la forma 13.2 se muestra una escala de actitud elaborada con este procedimiento. La calificación de una persona en esta escala es la suma de los valores numéricos (0, 1, 2, 3 o 4) de las respuestas que ella elige.

TABLA 13.1 Sugerencias para elaborar enunciados para una escala de actitud tipo Likert

1. Los enunciados deben referirse al presente más que al pasado.
2. Los enunciados no deben ser objetivos ni susceptibles de interpretarse como objetivos.
3. Los enunciados no deben interpretarse en más de un sentido.
4. Los enunciados deben ser relevantes para el concepto psicológico que se analiza.
5. Los enunciados deben ser oraciones sencillas que incluyan sólo un razonamiento y no oraciones compuestas ni complicadas.
6. Deben evitarse enunciados que contengan negaciones dobles, palabras quizá poco comprensibles para los participantes, palabras con más de un significado, adjetivos o adverbios no específicos (por ejemplo, muchos, en ocasiones) o universales (como todos, siempre, ninguno o nunca).
7. Debe evitarse usar coloquialismos o jergas, ya que tienden a volver ambiguas y poco claras las oraciones.

FORMA 13.2 Escala para medir actitudes hacia las matemáticas o la ciencia

Instrucciones: Escriba su nombre en la esquina superior derecha. Cada uno de los enunciados de este cuestionario de opiniones expresa un sentimiento o una actitud hacia las matemáticas (ciencia). Usted debe indicar, en una escala de cinco puntos, el grado de concordancia entre la actitud expresada en cada enunciado y su propia opinión. Los cinco puntos son Muy en Desacuerdo (MD), en Desacuerdo (D), Indeciso (I), de Acuerdo (A), Muy de Acuerdo (MA). Marque (✓) la(s) letra(s) que mejor indique(n) en qué medida está de acuerdo o en desacuerdo con la actitud expresada en cada enunciado como *usted lo percibe*.

	1	2	3	4	5
1. Matemáticas (ciencia) no es una materia muy interesante.	MD	D	I	A	MA
2. Quiero desarrollar mis habilidades matemáticas (ciencia) y estudiar más esta materia.	MD	D	I	A	MA
3. Matemáticas (ciencia) es una materia muy valiosa y necesaria.	MD	D	I	A	MA
4. Las matemáticas (ciencia) me hacen sentir nervioso e incómodo.	MD	D	I	A	MA
5. En general he disfrutado al estudiar matemáticas (ciencia) en la escuela.	MD	D	I	A	MA
6. No quiero tomar más cursos de matemáticas (ciencia) de los que inevitablemente tenga que tomar.	MD	D	I	A	MA
7. Otras materias son más importantes para las personas que las matemáticas (ciencia).	MD	D	I	A	MA
8. Me siento muy tranquilo y nada temeroso cuando estudio matemáticas (ciencia).	MD	D	I	A	MA
9. Casi nunca me ha gustado estudiar matemáticas (ciencia).	MD	D	I	A	MA
10. Estoy interesado en adquirir más conocimientos de matemáticas (ciencia).	MD	D	I	A	MA
11. Las matemáticas (ciencia) ayudan a desarrollar la mente y enseñan a las personas a pensar.	MD	D	I	A	MA
12. Las matemáticas (ciencia) me hacen sentir incómodo y confundido.	MD	D	I	A	MA
13. Las matemáticas (ciencia) son para mí disfrutables y estimulantes.	MD	D	I	A	MA
14. No estoy dispuesto a tomar más que la cantidad requerida de matemáticas (ciencia).	MD	D	I	A	MA
15. Las matemáticas (ciencia) no son especialmente importantes en la vida cotidiana.	MD	D	I	A	MA
16. Tratar de entender las matemáticas (ciencia) no me provoca angustia.	MD	D	I	A	MA
17. Las matemáticas (ciencia) no son aburridas.	MD	D	I	A	MA
18. Pienso tomar tantos cursos de matemáticas (ciencia) como sea posible durante mis estudios.	MD	D	I	A	MA
19. Las matemáticas (ciencia) han contribuido en gran medida al progreso de la civilización.	MD	D	I	A	MA
20. Matemáticas (ciencia) es una de mis materias más temidas.	MD	D	I	A	MA
21. Me gusta intentar resolver problemas nuevos en matemáticas (ciencia).	MD	D	I	A	MA
22. No estoy motivado para trabajar mucho en problemas matemáticos (científicos).	MD	D	I	A	MA
23. Matemáticas (ciencia) no es una de las materias más importantes que deba estudiar la gente.	MD	D	I	A	MA
24. No me molesta trabajar con problemas matemáticos (científicos).	MD	D	I	A	MA

(Basado en Aiken, 1979a. Reproducido con autorización de la School Science and Mathematics Association.)

No todas las escalas de actitud denominadas escalas Likert se elaboraron, de hecho, mediante procedimientos de análisis de reactivos. En muchos casos, simplemente se reúne como instrumento un conjunto de enunciados declarativos, cada uno con cinco categorías de respuesta de acuerdo-desacuerdo, sin ninguna construcción teórica en mente o sin seguir el procedimiento de Likert. En consecuencia, no podemos estar seguros de que un cuestionario que parece una escala Likert en verdad fue elaborado mediante el proceso de escalas de Likert.

A pesar del mal uso del método de rangos sumarizados, tiene varias ventajas sobre el método de intervalos de igual aparición (Selltiz, Wrightsman y Cook, 1976). Debido a que no se requiere de jueces expertos, sin sesgo, es más fácil elaborar una escala tipo Likert que una Thurstone. Además, a diferencia de las escalas tipo Thurstone, las Likert permiten el uso de reactivos que están claramente relacionados con la actitud evaluada, mientras estén correlacionados significativamente con las calificaciones totales. Por último, es probable que una escala Likert tenga mayor coeficiente de confiabilidad que una escala Thurstone con la misma cantidad de reactivos. Sin embargo, al igual que las escalas Thurstone, las Likert han sido criticadas por el hecho de que distintos patrones de respuestas pueden producir la misma calificación y porque, en el mejor de los casos, las calificaciones representan mediciones ordinales.

Escalas tipo Guttman. De menor popularidad que los procedimientos de Thurstone y de Likert, un tercer procedimiento de escalas de actitud es el *análisis de escalograma* desarrollado por Louis Guttman. El objetivo de un análisis de escalograma (Guttman, 1944) es determinar si las respuestas a los reactivos seleccionados para medir una actitud dada fallan en una dimensión única. Cuando los reactivos conforman una escala unidimensional verdadera, el participante que elige un reactivo en particular también está aceptando todos los reactivos que tienen un valor de escala menor. Es más probable que ocurra esta situación con los reactivos de pruebas cognoscitivas que con enunciados de actitud u otros reactivos sobre afectividad.

Como en el método de Bogardus (1925) sobre construcción de escalas de actitud, el objetivo del análisis de escalograma es producir una escala ordinal acumulativa. Guttman advirtió la dificultad de elaborar una escala de intervalos verdadera con reactivos de actitud, pero sintió que podría realizarse una aproximación. La medida en que se logra una verdadera escala está indicada por el *coeficiente de reproductibilidad*, calculado como la proporción de las respuestas reales que caen en el patrón perfecto de una escala Guttman verdadera. Es decir, ¿qué proporción de los participantes que aceptan un reactivo en particular aceptan todos los reactivos que están más abajo de él en la escala? Un valor aceptable del coeficiente de reproductibilidad está alrededor de .90.

En el cuadro 13.1 se presenta un ejemplo de matriz de respuesta para calcular el coeficiente de reproductibilidad para una escala Guttman de siete enunciados aplicada a 10 personas. Observe que los participantes (filas) se ordenan de acuerdo con la cantidad total de respuestas +, donde el signo de + indica acuerdo y el de − indica desacuerdo con la actitud expresada en el enunciado en particular. Después se cuenta la cantidad de respuestas + (p) para cada enunciado y se traza una línea divisoria bajo la fila que corresponde a la respuesta número p. Por ejemplo, hay nueve respuestas + en la columna para el enunciado 5, de modo que se traza una línea horizontal bajo la novena entrada de esa columna. En una escala Guttman acumulativa, todas las respuestas por arriba de esta línea deben ser + y todas las respuestas por abajo deben ser −. Debido a que la respuesta − en la fila F y la respuesta + en la fila G de la columna 5 son desviaciones de este patrón ideal, se cuentan como errores. En consecuencia, hay dos errores para el enunciado 5. El número de errores para los enunciados restantes se determina de manera similar, produciendo un total de $E = 8$ errores para los siete enunciados combinados. A continuación se compara el número total de respuestas como la cantidad de enunciados multiplicada por el número de respuestas como N = filas × columnas = $10 \times 7 = 70$. El coeficiente de reproductibilidad (R) se calcula entonces con la fórmula siguiente:

$$R = 1 - \frac{E}{N} \tag{13.1}$$

como

$$1 - \frac{8}{70} = .886.$$

Dado que el valor de *R* mínimo aceptable para una escala Guttman verdadera es .90, el coeficiente de reproductibilidad para esta escala de siete reactivos es una evidencia no concluyente de que estos enunciados conforman una escala Guttman.

Otros procedimientos de escalamiento de actitud

Varios procedimientos más se han aplicado al proceso de elaboración de escalas de actitud, incluyendo la técnica del diferencial semántico, la técnica Q, el cálculo de magnitud, el escalamiento de expectativa de valor, y el análisis de facetas. Los primeros dos de estos procedimientos se examinan en el capítulo 16; los últimos tres se analizan a continuación.

CUADRO 13.1

MATRIZ DE RESPUESTA PARA CALCULAR EL COEFICIENTE DE REPRODUCTIBILIDAD DE UNA ESCALA GUTTMAN ILUSTRATIVA DE SIETE REACTIVOS

	ENUNCIADO							
Participante	**3**	**1**	**7**	**6**	**5**	**4**	**2**	**Total (+)**
I	+	+	+	+	+	+	+	7
B	+	+	+	+	+	–	+	6
A	–	+	+	–	+	–	+	4
E	–	+	+	–	+	–	+	4
H	–	+	+	–	+	+	–	4
J	–	+	+	–	+	–	+	4
D	–	+	+	–	+	–	–	3
C	–	–	+	–	+	–	–	2
F	+	–	–	–	–	–	–	1
G	–	–	–	–	+	–	–	1
Total (+)	3	7	8	2	9	2	5	
Errores	2	0	0	0	2	2	2	

Cálculo de magnitud. En este método, que se basa en un procedimiento psicofísico para escalar intensidades de estímulos, el participante asigna un valor numérico a cada una de las series de estímulos que varían a lo largo de un rango de intensidades. Las respuestas de una muestra representativa de personas se promedian entonces y se calculan contra las intensidades de estímulos reales. Se ha empleado un procedimiento similar para escalar las percepciones de acontecimientos sociales y políticos, tales como la gravedad de ciertos actos criminales (Sellin y Wolfgang, 1964). Las clasificaciones promediadas otorgadas a cada uno de los acontecimientos se calculan contra una medida de valores reales, tales como el costo monetario de un delito. Esta técnica también se ha usado para escalar otras percepciones sociales o actitudes, como la popularidad de candidatos políticos, al hacer que los participantes dibujen una línea cuya longitud refleje la fuerza de sus sentimientos o actitudes hacia la persona, objeto o acontecimiento.

Escalamiento de valores de expectación. Este método del escalamiento de actitudes fue propuesto por Fishbein y Ajzen (1975). El participante empieza por indicar la medida en que aprueba una serie de dimensiones afectivas o de valor (el componente *afectivo* o de *valor*). Después se pide al participante que señale la medida en que considera que cada una de estas dimensiones se aplica al asunto considerado (el componente *cognoscitivo* o de *expectativa*). Combinar cada expectativa con su correspondiente valor proporciona una calificación E-V. Por ejemplo, una investigación descrita por Fishbein y Ajzen (1975) evaluaba las preferencias entre diversas tecnologías de energía (nuclear, de combustibles fósiles, fuerza de las mareas y otros). Los participantes empezaban indicando su nivel de agrado o desagrado para cada una de las siguientes dimensiones relevantes como una característica de cada tecnología: bajo costo, riesgo de catástrofe, contaminación a corto y largo plazo, y repercusiones tecnológicas. Después los participantes indicaron, en términos de una cifra de probabilidad, la medida en que cada una de estas dimensiones caracterizaba a las tecnologías. Se asociaron altas probabilidades a las dimensiones preferidas y bajas probabilidades a las dimensiones rechazadas. Así, una tecnología favorecida podría identificarse mediante las probabilidades altas y una tecnología no favorecida se identificaría por las bajas probabilidades asignadas a la mayoría o a todas las dimensiones.

Análisis de facetas. Una crítica al análisis de escalograma, que también se aplica a los métodos de intervalos de igual aparición y de rangos sumarizados, es que las actitudes son estados complejos, multidimensionales, que raramente pueden representarse con una única calificación. Otra crítica es que la dimensionalidad de una escala de actitud puede variar con la muestra de los participantes. En cualquier caso, la investigación subsecuente de Guttman sobre medición de actitudes, que llamó "análisis del espacio mínimo" o *análisis de facetas*, tiene poca semejanza con su anterior interés en el *análisis de escalograma*. El *análisis de facetas* es un paradigma previo, multidimensional para la elaboración y el análisis de reactivos que puede aplicarse a cualquier actitud, objeto o situación (Castro y Jordan, 1977). El procedimiento se ha empleado para elaborar escalas transculturales de actitud-comportamiento con respecto a varias condiciones y situaciones psicosociales, incluyendo retraso mental e interacción racial-étnica (Hamersma, Paige y Jordan, 1973).

Otros procedimientos multidimensionales. Durante las últimas dos o tres décadas, se ha vuelto cada vez más obvio que la medición de actitudes es, en sentido estricto, multidimensional, y que se requieren procedimientos de evaluación más complejos. El uso del análisis facto-

rial con instrumentos de medición de actitudes es actualmente bastante común. Hay una tendencia a abandonar las escalas unidimensionales, lo cual se evidencia en el creciente uso de métodos como las escalas multidimensionales, el análisis de estructura latente, el análisis de partición latente y la técnica de rejilla de repertorio en las escalas de actitudes (vea Ostrom, Bond, Krosnick y Sedikides, 1994; Procter, 1993).

Fuentes de escalas de actitud

Una gran variedad de inventarios y escalas para evaluar actitudes se describe en una serie de libros publicados por el Instituto de Investigación Social de la Universidad de Michigan (por ejemplo, Robinson, Shaver y Wrightsman, 1991, 1999). Otras fuentes incluyen el *American Social Attitudes Data Sourcebook, 1947-78* (Converse, Dotson, Hoag y McGee, 1980) y *A Sourcebook of Harris National Surveys*: *Repeated Questions, 1963-76* (Martin y McDuffee, 1981). También se incluyen docenas de mediciones de actitud en una amplia gama de áreas en *Tests in Microfiche* (Educational Testing Service) y en el volumen 5 del *ETS Test Collection Catalog* (1991) (páginas Web www.ets.org y ericae.net/testcol.htm). Otra fuente de información sobre inventarios, cuestionarios y escalas de actitudes y valores publicados son los sitios Web de los editores y distribuidores de instrumentos de evaluación psicológica (vea el apéndice C).

Entre las actitudes que se han evaluado y estudiado ampliamente en Estados Unidos se encuentran, en orden alfabético, las actitudes hacia las personas de edad (viejas), el sida, las computadoras, el Congreso, las guarderías, la bebida, el ambiente, los roles de género, los niños superdotados, las personas discapacitadas, los homosexuales y las lesbianas, las opiniones mayoritarias, las matemáticas, los políticos, las relaciones sexuales prematrimoniales, el presidente, la raza (grupos étnicos), la escuela, la ciencia, el sexo, el tabaquismo, los maestros, las pruebas, las mujeres y el trabajo. Algunos de los editores y distribuidores de los instrumentos de evaluación psicológica incluidos en el apéndice C ofrecen a la venta cuestionarios y escalas de actitud. Además de los instrumentos de medición de actitud ya preparados, están disponibles programas de cómputo para generar instrumentos personales propios de ciertas organizaciones empresariales. Un ejemplo es el Generador de Estudios de Actitud de Empleados Easy.Gen (Wonderlic Personnel Test, Inc.). Este paquete de cómputo diseña preguntas de actitud y selecciona otras a partir de una base de datos de 515 preguntas que abarcan 41 temas. También aplica cuestionarios de actitud y produce gráficas e informes de los resultados.

Confiabilidad y validez de las mediciones de actitud

Debido a la homogeneidad de su contenido, la consistencia interna de los coeficientes de confiabilidad de las calificaciones de escalas de actitud publicadas e inéditas a menudo están en el .80 o incluso en el .90. Las confiabilidades test-retest tienden a caer un poco más abajo, pero con todo son bastante altas para las escalas tipo Thurstone y Likert.

Además de los cambios efectivos en las actitudes que produce alguna situación manipulada, diversas variables de situación y procedimiento pueden afectar la confiabilidad de un instrumento de actitud. Entre éstas se encuentran situaciones de aplicación, número de categorías de respuesta y método de calificación. La estandarización de un instrumento psicométrico implica condiciones de aplicación estándares y uniformes. Sin embargo, a menudo es imposible mantener constantes las condiciones de aplicación cuando se recopilan datos sobre actitudes en distintos tipos de situaciones. Debido a que la confiabilidad implica consistencia de diferencia-

ción entre personas, la confiabilidad de un instrumento psicométrico tiende a ser inferior cuando las condiciones en que se aplica tienen efectos distintos en las calificaciones de distintas personas. Por ejemplo, las calificaciones de los niños más pequeños resultan más afectadas que las que corresponden a niños mayores por las variaciones en las condiciones de aplicación de las escalas de actitud. Por consiguiente, no es de sorprender que los coeficientes de confiabilidad obtenidos de las calificaciones de escalas de actitud aumenten con la edad cronológica de los participantes.

Otro factor que puede influir en la confiabilidad de una escala de actitud es la cantidad de categorías de respuesta. Las calificaciones de los instrumentos con un mayor número de categorías de respuesta de reactivos tienden a tener varianzas mayores y, por ende, confiabilidad más elevada que las calificaciones de instrumentos con cantidades menores de categorías de respuesta. Ésta es una de las posibles razones de que las calificaciones de las escalas tipo Likert con seis, nueve e incluso más categorías de respuestas no tengan coeficientes de confiabilidad perceptiblemente mayores que los correspondientes a las cinco categorías tradicionales. Cuando una escala de clasificación tiene una cantidad de categorías mayor, parece que los clasificadores son incapaces de realizar las discriminaciones precisas requeridas y, por lo tanto, sólo usan algunas de las categorías. Una posible excepción de esta regla ocurre cuando el rango de actitudes hacia un concepto específico es pequeño, en cuyo caso aumentar la cantidad de categorías de respuesta a seis o siete puede tener un pequeño efecto sobre la confiabilidad (Masters, 1974).

También se ha sugerido que aumentar el número de categorías de respuesta puede mejorar la confiabilidad general si las respuestas se transforman a calificaciones normales desviadas (z) (Wolins y Dickinson, 1973). Esta técnica es uno de los múltiples esfuerzos por aumentar la confiabilidad y la validez de las mediciones de actitud mediante cierto tipo de ponderación de reactivos o de componentes. Desafortunadamente, ninguno de los diversos esquemas de ponderación diferencial ha resultado ser superior con respecto a su confiabilidad hacia procedimientos de calificación más tradicionales. Esto es particularmente cierto cuando la cantidad de reactivos de un instrumento de calificación única es grande.

En ciertos tipos de reactivos de actitud, incluyendo algunas de las escalas tipo Thurstone, hay una categoría de respuesta neutral o intermedia (?, No sé, Inseguro u Otro), además de las categorías bipolares de Sí-No. El uso de esta categoría neutral por parte de los participantes varía de acuerdo con las instrucciones, el contexto y el tipo de objeto de actitud. No obstante, Alwin y Krosnick (1991) descubrieron que la confiabilidad de las mediciones de categorías múltiples de las actitudes sociopolíticas no aumentaba proporcionando explícitamente una opción para "no sé". Por otra parte, en formatos de dos y tres categorías, la inclusión de una categoría de respuesta neutral puede mejorar la confiabilidad en cierta medida (Aiken, 1983c). Por esta razón, en general se recomienda incluir una categoría de respuesta neutral en las escalas que constan de reactivos que se calificarán en forma dicotómica.

Con respecto a su validez, las calificaciones de las escalas de actitud tienden a hacer contribuciones pequeñas pero significativas a la predicción del desempeño en las materias escolares y los ambientes institucionales. Las mediciones de actitud, en general, no se han correlacionado en gran medida con el comportamiento real, y las reseñas de investigación han concluido que no pronostican con mucha precisión los comportamientos específicos. Sin embargo, Ajzen y Fishbein (1977) proporcionaron evidencia de que el comportamiento específico puede pronosticarse a partir de mediciones de actitud hacia el comportamiento específico, especialmente cuando los enunciados sobre actitudes se expresan en términos de conductas.

MEDICIÓN DE VALORES

Los *valores* que sostiene una persona —utilidad, importancia o mérito atribuido a actividades u objetos particulares– están relacionados con los intereses y las actitudes, pero no son idénticos a éstos. En comparación con las actitudes, los valores se consideran más importantes para la personalidad y más básicos para la expresión de necesidades y deseos individuales. El concepto de *valor* no se limita, desde luego, a la psicología y la sociología, sino que se emplea en filosofía, religión, economía y otros campos.

Las mediciones de valores se han usado en una amplia variedad de investigaciones en todo el mundo para comparar valores relacionados con el trabajo, la moral, la crianza de los niños, entre otros, en distintas culturas. Muchos de los antiguos instrumentos diseñados para medir valores, tales como el popular Estudio de Valores (Allport, Vernon y Lindzey, 1960), eran similares en contenido a los inventarios de intereses, actitudes y creencias. Muchos de estos instrumentos, que resultaron ser incompatibles con posteriores conceptos sobre valores (Braithwaite y Scott, 1991), ahora están descontinuados. El Estudio de Valores de Rokeach es una excepción.

Estudio de Valores de Rokeach

Milton Rokeach, quien dirigió una amplia investigacion internacional transcultural sobre el tema, definió el valor como "una organización de creencias relativamente perdurable alrededor de un objeto o situación, que predispone a la persona a reaccionar de determinada manera" (Rokeach, 1968, p. 112). Rokeach sostuvo que hay dos tipos de valores: aquellos relativos a los modos de conducta (*valores instrumentales*) y los concernientes a estados finales (*valores terminales*). Aunque en gran medida los psicólogos vocacionales han limitado su atención a los valores terminales, Rokeach definió varias subcategorías tanto de valores instrumentales como de terminales y diseñó un instrumento para medirlos.

Rokeach clasificó los valores instrumentales en dos tipos: *valores morales* y *valores de competencia*. La primera categoría se refiere a las formas de conducta interpersonal, las cuales producen sentimientos de culpa cuando se violan. La segunda categoría, los valores de competencia, tiene que ver con formas de conducta intrapersonales, de autorrealización, cuya violación provoca sentimientos de inadecuación. Los valores terminales también se subdividen en *valores personales* y *valores sociales*.[1] Los valores personales, que incluyen estados finales como la conciencia tranquila y la salvación, están centrados en la persona. Los valores sociales, que incluyen estados finales como la equidad y la paz mundial, se centran en la sociedad.

El Estudio de Valores de Rokeach (de CPP) consta de una serie de 18 términos o frases de valores instrumentales y 18 de valores terminales para evaluar la importancia relativa que tienen estos valores para las personas. Se instruye al participante que ordene los 18 reactivos de cada lista en orden de importancia para él. Ningún otro instrumento intenta evaluar tantos valores, un hecho que, aunado a la velocidad de administración y de calificación, así como al reducido costo, ha contribuido a su popularidad. El Estudio de Valores de Rokeach tiene una confiabilidad

[1]La distinción entre personal y social también se incluyó en una definición posterior de valor como "una creencia perdurable de que una forma de conducta o estado final de existencia en particular es personal o socialmente preferible a una forma de conducta o estado final de existencia opuesto o inverso" (Rokeach, 1973, p. 5).

adecuada para diferenciar entre grupos, un propósito para el que se ha empleado en cientos de investigaciones durante tres décadas. Personas de distintas nacionalidades y ubicadas en diferentes áreas de la vida ordenan los reactivos del Estudio de Valores de Rokeach de manera distinta. Por ejemplo, los estudiantes israelitas asignaron las clasificaciones mayores a "un mundo en paz" y "seguridad nacional", mientras que los estudiantes estadounidenses otorgaron un valor superior a "una vida cómoda" y "ambición" (Rokeach, 1973, 1979). En otro estudio transcultural de los sistemas de valor, los estudiantes universitarios de Australia otorgaron sitios significativamente más elevados que los estudiantes universitarios chinos a los siguientes valores del estudio de Rokeach: una vida excitante, un mundo en paz, seguridad para la familia, felicidad, armonía interior, estar alegre, perdonar, ayudar, ser honesto, ser amoroso y ser responsable. En contraste, los estudiantes chinos asignaron lugares significativamente más altos que los australianos a un mundo de belleza, la seguridad nacional, el placer, el reconocimiento social, la sabiduría, ser ambicioso, ser capaz, ser valeroso, ser imaginativo, ser intelectual, ser lógico y ser autocontrolado (Feather, 1986).

Valores vocacionales

Un cambio que ha ocurrido en años recientes en la medición de valores es la inclusión de tales mediciones en inventarios de base amplia orientados a la consejería vocacional y a las actitudes o motivaciones hacia el trabajo. Sin embargo, los títulos de varios instrumentos publicados, diseñados específicamente para evaluar elecciones y satisfactores ocupacionales aún contienen el término "valores". Entre éstos se encuentra el Inventario de Valores para el Trabajo y la Escala de Valores. Los valores vocacionales medidos por estos intrumentos varían de persona a persona, en la misma persona de un momento a otro, y según el carácter del trabajo. Super (1973) encontró, por ejemplo, que las personas de ocupaciones de nivel superior estaban más motivadas por la necesidad de autorrealización, que es un objetivo intrínseco, mientras que los valores extrínsecos era más probable que estuviesen suscritos por personas de ocupaciones de nivel inferior.

La Escala de Valores. Desarrollada por Work Importance Study, un consorcio internacional de psicólogos vocacionales de los Estados Unidos de Norteamérica, Asia y Europa, la Escala de Valores (de CPP) posee características tanto del Inventario de Valores para el Trabajo como del Estudio de Valores de Rokeach. El propósito de este consorcio y de la Escala de Valores fue comprender los valores que los individuos buscan o esperan encontrar en diversos papeles de la vida y evaluar la importancia relativa del papel del trabajo como medio de realización de valores en el contexto de otros papeles en la vida. La Escala de Valores consta de 106 reactivos, toma de 30 a 45 minutos completarla, y se califica para 21 valores (cinco reactivos por valor):

Utilización de habilidades	Creatividad	Interacción Social
Aprovechamiento	Recompensas económicas	Relaciones Sociales
Progreso	Estilo de vida	Variedad
Estética	Desarrollo personal	Condiciones de trabajo
Altruismo	Actividad física	Identidad cultural
Autoridad	Prestigio	Destreza física
Autonomía	Riesgo	Seguridad económica

Las confiabilidades de todas las escalas son adecuadas para la evaluación individual en el nivel adulto; las confiabilidades de diez escalas son lo bastante elevadas como para permitir la evaluación individual en el nivel universitario, y las confiabilidades de ocho escalas son adecuadas a nivel de bachillerato. En el manual se proporcionan las medias y las desviaciones estándar para tres muestras (bachillerato, universidad, adultos) así como los datos de la validez de constructo del intrumento.

ORIENTACIONES PERSONALES

Las mediciones de *orientaciones personales* son similares a los inventarios de intereses y de valores. Un ejemplo de orientación personal es la medida en que un individuo intenta "ser todo lo que puede ser" —para alcanzar su potencial o *autorrealizarse*. Otros conceptos relativos a la orientación personal sobre los que se han diseñado medidas y realizado investigaciones son orientación de vida (Dudek y Makowska, 1993; Madhere, 1993), orientación moral (Lidell, Halpin y Halpin, 1992), orientación interpersonal (Silva, Martinez, Moro y Ortet, 1996) y rol sexual (género).

Inventario Bem sobre el Papel del Sexo

El desarrollo de varias mediciones relativas al papel del sexo durante los últimos 30 años, aproximadamente, fue impulsado en gran medida por el interés sobre discriminación de género y por el carácter y origen de las diferencias de sexo en las características psicológicas. Uno de los instrumentos de este tipo más prominentes es el Inventario Bem sobre el Papel del Sexo (BSRI) (Bem, 1974; publicado por Mind Garden). La forma breve de este inventario (BSRI breve), que se diseñó para clasificar a las personas de acuerdo con su orientación en cuanto al papel del sexo, consiste en 60 palabras o frases que deben clasificarse en una escala de siete puntos donde 1 significa *nunca o casi nunca es cierto* y 7 se refiere a *siempre o casi siempre es cierto*. 20 de estos reactivos corresponden a características consideradas significativamente más deseables en hombres que en mujeres (por ejemplo, agresivo, ambicioso), 20 reactivos se refieren a características que se cree son considerablemente más adecuadas para mujeres que para hombres (por ejemplo, afectivo, alegre), y 20 reactivos supuestamente son neutrales sobre el sexo (por ejemplo, adaptable, minucioso). Primero se determinan las calificaciones en tres escalas, Masculinidad (M), Feminidad (F) y Androginia (A). Después, el examinado se ubica en una de las siguientes cuatro categorías de acuerdo con sus calificaciones en las escalas M, F y A. Masculino (superior a la mediana en M e inferior a la mediana en F), Femenino (superior a la mediana en F e inferior a la mediana en M), Andrógino (superior a la mediana tanto en M como en F) o Indiferenciado (inferior a la mediana tanto en M como en F).

Las confiabilidades test-retest y de consistencia interna del BSRI breve son, en general, satisfactorias, pero los datos de validez son bastante escasos. Asimismo, se advierte a los investigadores que no confíen en las normas incluidas en el manual: sólo están basadas en muestras de los estudiantes de la Universidad de Stanford.

Escala de Igualitarismo del Papel de los Sexos

Otro inventario relativo a diferencias psicológicas y de comportamiento entre los sexos es la Escala de Igualitarismo del Papel de los Sexos (SRES) (por L. A. King y D. W. King; Sigma As-

sessment Systems). Diseñada para medir las actitudes hacia la igualdad de hombres y mujeres, la SRES contiene 95 enunciados de actitud que se responden en escalas de cinco puntos. Los reactivos están escritos en un nivel de comprensión de 6° o 7° grados y requieren de aproximadamente 25 minutos para contestarse. Las escalas de 19 reactivos de la SRES cubren los siguientes dominios: Papeles maritales, Papeles de los padres, Papeles de empleos, Papeles sociales interpersonales heterosexuales, y Papeles educativos.

Las confiabilidades de consistencia interna, de test-retest y de formas alternas de la SRES son bastante altas. Con respecto a su validez, las relaciones entre la SRES y otras variables, así como las diferencias de grupo en las calificaciones, atestiguan la validez convergente, discriminante y de constructo de este instrumento (King y King, 1993).

Inventario de Orientación Personal

Este inventario de amplio uso (por E. L. Shostrom; EdITS) fue diseñado para medir valores y comportamientos importantes en el desarrollo de las personas autorrealizadas. Dicha gente desarrolla y usa todo su potencial y está libre de las inhibiciones y agitación emocional que caracteriza a las personas menos autorrealizadas. El Inventario de Orientación Personal (POI) consta de 150 reactivos de elección forzada adecuados para estudiantes de bachillerato, universitarios y adultos. Se califica primero en dos proporciones de orientación principales: Proporción de tiempo (Incompetencia de tiempo/Competencia de tiempo) y Proporción de apoyo (Apoyo externo/Apoyo interno). La Proporción de tiempo indica si la orientación de tiempo del participante está sobre todo en el presente, el pasado o el futuro. La Proporción de apoyo indica si la orientación de reactividad del participante es básicamente hacia otras personas o hacia sí mismo. Después de que estas dos proporciones se han calculado, se determinan las calificaciones de las siguientes diez escalas:

Valor de autorrealización. Afirmación de los valores primarios de las personas autorrealizadas.

Existencialidad. Capacidad de reaccionar de manera situacional o existencial.

Reactividad de sentimientos. Sensibilidad de reacción a los propios sentimientos y necesidades.

Espontaneidad. Libertad de actuar espontáneamente para ser uno mismo.

Autorrespeto. Afirmación de sí mismo debido al esfuerzo.

Autoaceptación. Afirmación o aceptación de sí mismo a pesar de las debilidades o deficiencias.

Naturaleza del hombre. Grado de visión constructiva de la naturaleza del hombre, la masculinidad y la feminidad.

Sinergia. Capacidad de ser sinérgico, de trascender dicotomías.

Aceptación de la agresión. Capacidad de aceptar la agresividad natural de uno mismo como opuesta a la defensividad, la negación y la represión de la agresión.

Capacidad de contacto íntimo. Capacidad de desarrollar relaciones íntimas con otros seres humanos, libre de expectativas y obligaciones.

En el manual se incluyen las normas por rangos percentilares basadas en una muestra de 2,607 estudiantes universitarios de primer año (1,514 hombres y 1,093 mujeres), además de califica-

ciones promedio y perfiles basados en una muestra de adultos menor. Las confiabilidades test-retest de las escalas individuales, calculadas en una muestra de 48 estudiantes universitarios, son moderadas (mayormente de .60 y .70). También se proporcionan en el manual las correlaciones con otras escalas de personalidad y otra evidencia de la validez del POI.

Inventario de Orientación de Vida

Este inventario (Udai, 1995) se usa en contextos institucionales y profesionales para el desarrollo de recursos humanos y propósitos de investigación. Permite a los solicitantes, empleados o asesorados, evaluar sus orientaciones de estilo de vida con respecto a los conceptos de *incrementar* o *envolver*. Un estilo de vida de incremento está orientado a la innovación, el cambio y el crecimiento, mientras que un estilo de vida envolvente se orienta hacia las metas de la estabilidad tradicional y la fuerza interior. Hay dos escalas, la A y la B. La Escala A consiste en actividades de 14 reactivos de actividades correspondientes a las orientaciones de incremento y envolvente. El participante indica en una escala de 5 puntos la cantidad de tiempo que dedica a la actividad. En los seis pares de reactivos de elección forzada de la Escala B, el participante indica la importancia que cada una de las dos actividades tiene para él.

RESUMEN

Las actitudes son predisposiciones aprendidas para responder de manera positiva o negativa ante cierto objeto, determinada persona o situación. Como tales, hay características de personalidad, aunque en un nivel más superficial que el temperamento o los rasgos. Las actitudes pueden evaluarse de varias maneras, de las cuales las más populares son los inventarios o escalas de actitud. Los procedimientos para elaborar escalas de actitud fueron diseñados por Thurstone (método de intervalos de igual aparición), Likert (método de rangos sumarizados) y Guttman (análisis de escalograma). Otras técnicas de escalamiento de actitud incluyen el diferencial semántico, los tipos Q, el cálculo de magnitud, el escalamiento de valores de expectación, el análisis de facetas y diversos procedimientos estadísticos multivariados.

A juzgar por la diversidad de instrumentos disponibles, las personas pueden tener una actitud hacia casi cualquier cosa, por ejemplo, cualesquier tema escolar, vocación, grupo definido, institución, acción social propuesta, o práctica definida. La gran mayoría de los cientos de escalas y cuestionarios de actitud mencionados en diversas fuentes de referencia no está estandarizada, por haber sido diseñada para investigaciones o aplicaciones particulares. Sin embargo, muchos instrumentos estandarizados para evaluar actitudes hacia la escuela y las materias escolares, el trabajo y los jefes en el trabajo, y otros tipos de actividades humanas, están disponibles por parte de los editores de pruebas comerciales.

Los valores, o creencias, relativos a la utilidad o mérito de algo pueden evaluarse mediante muchos inventarios distintos. El Estudio de Valores de Rokeach y la teoría relacionada han estimulado la investigación sobre valores político-ideológicos y los conceptos sobre la buena vida. La investigación sobre valores relacionados con el trabajo ha originado varios instrumentos psicométricos, en especial el Inventario de Valores para el Trabajo y la Escala de Valores.

La creencia o confianza en la verdad o existencia de algo no es inmediatamente susceptible de someterse a una prueba rigurosa. Las creencias se basan en menos evidencia y se sostienen con mayor tenacidad que las opiniones y las actitudes. Las mediciones de creencias religiosas, morales, políticas y muchas otras se han diseñado con propósitos de investigación, pero la mayoría son *ad hoc* y no están estandarizadas.

Las orientaciones personales, tales como la autorrealización y la identificación de los papeles de los sexos, cruzan un rango de variables de personalidad un poco más amplio que las actitudes y los valores. El Inventario Bem sobre el Papel del Sexo se ha usado en numerosas investigaciones para diferenciar entre los papeles masculino, femenino y andrógino. Los resultados de la investigación con este inventario y con instrumentos similares indican que la masculinidad y la feminidad son dos conceptos psicológicos distintos y no simplemente opuestos polares en una dimensión única. En este capítulo se examinaron otras tres mediciones de multiescala para las orientaciones personales: la Escala de Igualitarismo del Papel de los Sexos, el Inventario de Orientación Personal y el Inventario de Orientación de Vida. El primero de estos instrumentos mide las actitudes hacia la igualdad de hombres y mujeres, el segundo mide diversos aspectos de la autorrealización, y el tercero evalúa la orientación del estilo de vida en términos de procesos de incremento o de envolvimiento.

PREGUNTAS Y ACTIVIDADES

1. En el proceso de construir una escala de actitud mediante el método de Thurstone de intervalos de igual aparición, cada uno de los 50 jueces distribuye 200 enunciados de actitudes en 11 pilas. Los números de los jueces que colocan los enunciados D, N y X en cada una de las 11 categorías aparecen en las tres distribuciones de frecuencia incluidas abajo. Calcule el valor de escala (mediana) y el índice de ambigüedad (rango semiintercuartilar) de cada enunciado mediante los métodos descritos en el apéndice A. Use el número de pila (1, 2, ..., 11) más .5 como el límite exacto superior del intervalo.

NÚMERO DE PILA	ENUNCIADO D	ENUNCIADO N	ENUNCIADO X
1			8
2			17
3		6	10
4		10	9
5		13	6
6	3	8	
7	7	6	
8	9	4	
9	13	3	
10	10		
11	8		

2. Copie los enunciados de la forma 13.1 (página 297) en orden de sus números de reactivo, asígneles un nuevo número del 1 al 12 y omita los valores de escala. A continuación realice múltiples copias de esta escala de actitud y aplíquela a distintas personas. Determine la calificación total de la escala de cada participante sumando los valores de escala de los enunciados que eligió y dividiendo la suma entre el número total de enunciados (12). Pida a los participantes que expliquen las razones de sus actitudes hacia la pena de muerte y resuma los resultados. ¿Qué variables de personalidad considera que están relacionadas con las actitudes hacia la pena de muerte?

3. Realice varias copias de la Escala de Actitud hacia las Matemáticas o la Ciencia de la forma 13.2 y aplíquela a varias personas. Calcule sus calificaciones en las cuatro partes de la escala: D (Disfrute de las matemáticas o la ciencia), M (Motivación en matemáticas o ciencia), I (Importancia de las

matemáticas o la ciencia) y T (Temor hacia las matemáticas o la ciencia). También determine una calificación total sumando las calificaciones D, M, I y T. La calificación D consiste en la suma de respuestas a los reactivos 1, 5, 9, 13, 17 y 21; la calificación M es la suma de respuestas a los reactivos 2, 6, 10, 14, 18 y 22; la calificación I es la suma de respuestas a los reactivos 3, 7, 11, 15 19 y 23; y la calificación T es la suma de respuestas a los reactivos 4, 8, 12, 16, 20 y 24. Las respuestas a los reactivos 1, 4, 6, 7, 9, 12, 14, 15, 17, 20, 22 y 23 se califican como SD = 4, D = 3, U = 2, A = 1, y SA = 0; las respuestas a los reactivos 2, 3, 5, 8, 10, 11, 13, 16, 18, 19, 21 y 24 se califican como SD = 0, D = 1, U = 2, A = 3, y SA = 4. La calificación T (total) es la suma de las calificaciones de las cuatro partes (T = D + M + I + T). Las calificaciones altas para T (total), D, M, I o T indican actitudes favorables hacia las matemáticas (o la ciencia), y las calificaciones bajas señalan actitudes desfavorables hacia estas materias. Pregunte a los participantes sobre las causas de sus actitudes hacia las matemáticas y la ciencia, y resuma sus hallazgos.

4. Usando el procedimiento descrito en el cuadro 13.1, calcule el coeficiente de reproductibilidad para los siguientes datos. Las filas de la matriz tendrán que reordenarse de manera adecuada antes de realizar los cálculos.

	ENUNCIADOS DE ACTITUD					
Participante	*1*	*2*	*3*	*4*	*5*	*6*
A	+	+	+	+	+	+
B	–	–	+	+	+	–
C	–	–	–	–	+	–
D	+	+	+	+	+	–
E	–	–	+	+	+	+
F	–	–	+	–	+	+
G	–	–	+	+	–	–

5. ¿De qué manera la orientación del papel del sexo está relacionada con la orientación sexual? ¿Esperaría usted que los hombres con una calificación alta en la escala de feminidad en un inventario de interés o de personalidad sean homosexuales, o que las mujeres con una calificación alta en la escala de masculinidad sean lesbianas? ¿Por qué sí o por qué no?

6. Durante los últimos años, los términos *valores familiares* y *valores cristianos* se han usado ampliamente por los políticos, los reformadores sociales y los medios de comunicación en general. ¿Qué son los valores familiares y de qué manera son distintos a los valores cristianos? Haga el intento de elaborar una escala consistente en un conjunto de reactivos para medir valores familiares. ¿Qué tipo de personas esperaría que obtuvieran una calificación alta y cuáles tendrían una calificación baja en su escala?

7. Aplique el Inventario de Valores Educativos (EVI) de la forma 13.3 a varios estudiantes de distinto origen y calcule sus calificaciones. Elabore y compare los perfiles de calificaciones de los estudiantes en las seis escalas de valores. Las respuestas a cada reactivo del EVI se califican en una escala de 1 a 5 de izquierda (N) a derecha (E), respectivamente. La suma de las calificaciones de los reactivos 2, 7, 18 y 21 es la calificación de Valor estético. La suma de las calificaciones de los reactivos 1, 8, 13 y 24 es la calificación del Valor de liderazgo. La suma de las calificaciones de los reactivos 5, 12, 16 y 19 es la calificación de Valor social. La suma de las calificaciones de los reactivos 6, 11, 17 y 22 es la calificación del Valor científico. La suma de las calificaciones de los reactivos 3, 10, 14 y 20 es la calificación del Valor vocacional.

FORMA 13.3 Inventario de Valores Educativos

Parte I. Cada uno de los reactivos de esta sección se refiere a un posible objetivo o énfasis de educación superior. Marque la letra adecuada después de cada uno de los siguientes enunciados para indicar qué tan importante cree que debería ser el objetivo correspondiente. Use estas claves de respuesta: N = No importante, A = Algo importante, I = Importante, M = Muy importante, E = Extremadamente importante.

1.	Capacidad para guiar o dirigir a otras personas.	N	A	I	M	E
2.	Aprecio por las cosas bellas y armoniosas de la vida.	N	A	I	M	E
3.	Preparación para la vocación o profesión elegida.	N	A	I	M	E
4.	Profundizar sobre el significado y propósito de la vida.	N	A	I	M	E
5.	Comprender los problemas sociales y sus posibles soluciones.	N	A	I	M	E
6.	Comprender las teorías científicas y las leyes de la naturaleza.	N	A	I	M	E
7.	Adquirir la capacidad de expresarse de manera artística.	N	A	I	M	E
8.	Comprender cómo dirigir a otras personas para el cumplimiento del mismo fin.	N	A	I	M	E
9.	Desarrollo de una filosofía personal en la vida.	N	A	I	M	E
10.	Aprender cómo tener éxito en la ocupación o campo elegidos.	N	A	I	M	E
11.	Aprender sobre problemas científicos y sus soluciones.	N	A	I	M	E
12.	Comprender a las personas de distintas clases sociales y culturas.	N	A	I	M	E

Parte II. Marque la letra adecuada después de los siguientes reactivos para calcular qué tan valiosos son los tipos de cursos universitarios para los estudiantes en general. Use estas claves: N = Nada valiosos, A = Algo valiosos, V = Valiosos, B = Bastante valiosos, M = Muy valiosos.

13.	Cursos relativos a cómo dirigir y organizar a las personas.	N	A	V	B	M
14.	Cursos en una vocación elegida o campo profesional.	N	A	V	B	M
15.	Cursos sobre ideas filosóficas y/o religiosas.	N	A	V	B	M
16.	Cursos relativos a la comprensión y ayuda a las personas.	N	A	V	B	M
17.	Cursos en ciencia y matemáticas.	N	A	V	B	M
18.	Cursos sobre música, arte y literatura.	N	A	V	B	M

Parte III. Marque la letra adecuada después de cada uno de los reactivos para indicar cuánta atención considera que debe darse a cada tipo de curso universitario en la educación de la mayoría de los estudiantes. Utilice estas claves: N = Nada de atención, P = Poca atención, M = atención Moderada, S = atención Superior al promedio, G = Gran cantidad de atención.

19.	Cursos relativos a cómo entender y ayudar a otras personas.	N	M	P	S	G
20.	Cursos en el campo vocacional o profesional de su elección.	N	M	P	S	G
21.	Cursos en arte, literatura y música.	N	M	P	S	G
22.	Cursos sobre campos científicos y matemáticos.	N	M	P	S	G
23.	Cursos sobre filosofía y religión.	N	M	P	S	G
24.	Cursos relativos a la organización y dirección de personas.	N	M	P	S	G

CAPÍTULO CATORCE

EVALUACIÓN DE LA PERSONALIDAD: ORÍGENES, APLICACIONES Y PROBLEMAS

El término *personalidad* tiene muchos significados. Para algunos se refiere a un carisma misterioso poseído por las estrellas de Hollywood y por otras personas populares e influyentes, pero no por cualquiera. Para otros, personalidad es lo mismo que temperamento —una predisposición natural, basada en lo genético, para pensar, sentir y actuar de cierta manera—. Todavía para otros, la personalidad consiste en la mezcla única que una persona tiene de rasgos emocionales, intelectuales y de carácter (honestidad, valor, etc.). Para los psicólogos de orientación más conductual, la personalidad no es algo interno, sino más bien un patrón observable de conducta organizada que es típico de una persona.

Quizá una convención aceptable sea definir la *personalidad* humana como un compuesto de habilidades cognoscitivas, intereses, actitudes, temperamento y otras diferencias individuales en los pensamientos, sentimientos y la conducta. Esta definición enfatiza el hecho de que la personalidad es una combinación única de características cognoscitivas y afectivas que puede describirse en términos de un patrón típico y bastante consistente de conducta individual.

A partir de la última definición se desprende que los métodos para evaluar la personalidad deberían incluir una gama amplia de variables cognoscitivas y afectivas. Entre esas variables se encuentran las medidas de aprovechamiento, inteligencia, habilidades especiales, intereses, actitudes y valores analizadas en los capítulos 6 a 13. Otras características emocionales, de temperamento y de estilo, a las que por tradición se ha denominado *variables de personalidad,* también son importantes en la comprensión y predicción de la conducta humana. En este capítulo se presentan material antecedente y aplicaciones que conciernen a la evaluación de la personalidad. En los cuatro capítulos siguientes se presentan métodos más específicos para evaluar la personalidad: observaciones, entrevistas, calificaciones, inventarios de personalidad y técnicas proyectivas.

PSEUDOCIENCIAS Y OTROS ANTECEDENTES HISTÓRICOS

Como sucedió con las pruebas de inteligencia, la evaluación de la personalidad se desarrolló en parte desde la investigación sobre las diferencias individuales y de grupo. Muchos antecedentes de la evaluación contemporánea de la personalidad pueden encontrarse en la historia de la psicología anormal y la psiquiatría. La historia de la ciencia está repleta de ejemplos de creencias o

doctrinas que en una época tuvieron muchos seguidores, pero que luego se demostró eran parcial o totalmente incorrectas. Entre esas doctrinas pseudocientíficas se encuentran la frenología, la fisionomía y la grafología.

La *frenología*, en la cual muy poca gente cree en la actualidad, era vista con seriedad por muchos hombres y académicos famosos a finales del siglo XVIII y principios del XIX. Según los defensores de la frenología, el desarrollo de áreas específicas en el cerebro humano está asociado con ciertas características de personalidad y trastornos mentales. Entre los rasgos de personalidad que se suponía estaban relacionados con protuberancias localizadas en ciertas regiones del cerebro están la afabilidad, la combatividad y la codicia (figura 14.1). Una consecuencia natural de esta creencia es que la personalidad puede ser analizada tocando la cabeza de una persona a fin de localizar "protuberancias" en áreas específicas del cerebro que se presume están asociadas con ciertas características. La frenología tuvo gran influencia en la psiquiatría del siglo XIX y en la *psicología de las facultades*, una noción que afectó los programas académicos de la época. De acuerdo con la psicología de las facultades, la mente consiste en una serie de facultades que pueden ser desarrolladas mediante el ejercicio mental (por ejemplo, estudiando latín,

FIGURA 14.1 Tabla de frenólogos de las funciones cerebrales. Localización de varias facultades afectivas y cognoscitivas descritas por un frenólogo del siglo XIX. La frenología fue una pseudociencia, y en la actualidad no tiene mucho crédito.

(Corbis)

griego, geometría y otras materias difíciles) de la misma manera que el cuerpo puede ser desarrollado por el ejercicio físico.

La *fisionomía*, otra pseudociencia, se interesa en determinar el temperamento y el carácter a partir de rasgos externos del cuerpo y en especial del rostro. Es posible advertir vestigios de la fisionomía en la selección del personal y los procedimientos de evaluación contemporáneos, por ejemplo, en el requisito de que una fotografía del solicitante acompañe a la solicitud de empleo. Otro instrumento de evaluación de la personalidad asociado con la fisionomía es la Prueba Szondi. Esta prueba consta de seis grupos de fotografías, cada grupo con ocho fotografías, de pacientes mentales con diferentes diagnósticos (por ejemplo, histeria, catatonia, paranoia, depresión o manía). En cada grupo los examinados seleccionan las fotografías que más les gustan y las que más les disgustan. La suposición básica que subyace a la Prueba Szondi es que los rasgos faciales de los pacientes mentales mostrados en las 12 fotografías seleccionadas y las 12 rechazadas tienen un significado personal para quien responde. Se supone que las necesidades y la personalidad del sujeto son similares a las de los pacientes mostrados en las fotografías. Como no se ha encontrado evidencia consistente que apoye la validez de la Prueba Szondi en el análisis de la personalidad o el diagnóstico psiquiátrico, la prueba ha sido desacreditada.

La creencia en la *grafología*, actividad donde se analiza la personalidad mediante el estudio de muestras de escritura, está quizá más difundida que la creencia en la fisionomía. Aunque tiene sentido suponer que la escritura, que es un tipo de conducta estilística, pueda reflejar características de personalidad, ni siquiera los analistas experimentados de la escritura se conocen por la precisión de sus interpretaciones. La fisionomía y la grafología tienen una mejor reputación que la frenología, pero muchas de las afirmaciones de sus defensores están igual de equivocadas.

No todos los intentos previos al siglo XX por desarrollar una ciencia de la evaluación de la personalidad deben etiquetarse como pseudociencia. Los esfuerzos de Francis Galton, Emil Kraepelin y Alfred Binet fueron muy respetables, aunque no siempre tuvieron éxito. En 1884, Galton propuso medir las emociones registrando cambios en el latido cardiaco y la tasa del pulso, y evaluar el buen humor, el optimismo y otros rasgos de personalidad observando a la gente en situaciones sociales inventadas. Kraepelin, quien es mejor conocido por su sistema de clasificación de los trastornos mentales, desarrolló la *técnica de asociación de palabras* en 1892. También durante la década de 1890, Alfred Binet, cuyo nombre el lector recordará de los capítulos sobre las pruebas de inteligencia, desarrolló métodos para estudiar las características de personalidad de las personas eminentes.

A pesar de unos inicios promisorios en el siglo XIX, un progreso genuino en la evaluación de la personalidad no llegó sino hasta el siglo XX. A este respecto, llaman particularmente la atención las pruebas de asociación de palabras de Carl Jung para detectar y analizar los complejos mentales (1905), la Hoja de Datos Personales de Robert Woodworth, el primer inventario estandarizado de personalidad que se aplicó de manera masiva (1919) y la Prueba de Manchas de Tinta de Hermann Rorschach (1920).

TEORÍAS DE LA PERSONALIDAD

Casi todos tenemos alguna teoría de por qué la gente se comporta como lo hace. Esas teorías de la naturaleza y la conducta humanas consisten, por lo común, en generalizaciones excesivas o estereotipos, pero sirven como guías aproximadas a las expectativas y la acción. En ocasiones la misma supervivencia de una persona depende de su capacidad para entender y predecir la conducta de otra gente.

Al percatarse de que todos somos diferentes de los demás y que la conducta humana puede ser muy compleja, los teóricos de la personalidad han aprendido a mostrarse suspicaces ante las explicaciones de sentido común. Ciertos psicólogos, impresionados por la individualidad y lo intrincado de las acciones humanas, han abandonado la esperanza de encontrar principios o leyes generales para explicar la personalidad. Rechazan el *enfoque nomotético*, la búsqueda de leyes generales de la conducta y la personalidad, como irreal e inadecuado para la tarea de comprender al individuo. En lugar de ello, abogan por un *enfoque idiográfico* de considerar a cada personalidad como un sistema legal, integrado y digno de estudio por derecho propio (Allport, 1937).

Existen muchas otras diferencias entre las teorías de la personalidad, siendo una el énfasis relativo que se pone en la herencia y el ambiente como moldeadores de la conducta. Los teóricos también difieren en el grado en que enfatizan las características internas individuales, o rasgos, más que las variables situacionales, como determinantes de la conducta. Como sugieren esos y otros puntos de debate entre los psicólogos, no existe una teoría de la personalidad que goce de aceptación general. Por el contrario, continuamente emergen y se modifican teorías y hallazgos de la investigación concernientes a los orígenes, la estructura y la dinámica de la personalidad. Con todo, para cualquiera que esté interesado en la evaluación psicológica, es importante estar al tanto de las diversas teorías de la personalidad y mostrarse escéptico ante las que no hayan sido probadas. A pesar de sus limitaciones, las teorías pueden servir como motivadores y guías en la medición y comprensión de la personalidad. Las teorías proporcionan marcos de referencia —ideas concernientes a la dinámica y el desarrollo de la personalidad y la conducta— para la interpretación de los hallazgos de la investigación. A este respecto, se supone que las teorías propuestas y probadas por los psicólogos profesionales son más útiles que las del sentido común.

Teorías de los tipos

Uno de los enfoques más antiguos para la comprensión de la personalidad es la noción de categorías o tipos fijos de gente. Galeno, un médico que vivió en la antigua Roma y estaba de acuerdo con la doctrina de Hipócrates de cuatro humores corporales (sangre, bilis amarilla, bilis negra y flema), sostenía que existen cuatro tipos de temperamento correspondientes. Se decía que el *tipo sanguíneo*, con un exceso de sangre, era vigoroso y atlético; el *tipo colérico*, con un exceso de bilis amarilla, se enfurecía con facilidad; el *tipo melancólico*, con un exceso de bilis negra, era por lo general depresivo o triste, y el *tipo flemático*, con un exceso de flema, se sentía crónicamente cansado o perezoso. Al igual que la frenología y otras nociones pseudocientíficas, la teoría humoral en la actualidad es sólo de interés histórico (pero vea la figura 14.3). Las teorías de los tipos corporales de Ernest Kretschmer, Cesare Lombroso y William Seldon, se basan con algo más de seguridad en datos observacionales, pero siguen siendo muy tentativas y generalizadas en exceso.

La idea de que la personalidad se asocia con el físico ha intrigado a muchos filósofos y poetas. Shakespeare lo declaró en muchas de sus obras. Por ejemplo, en el Acto I, Escena II de *Julio César*, César dice:

> Permítanme que me rodee de hombres robustos.
> Hombres de ceño liso, y sin preocupaciones.
> Yond Cassius tiene una apariencia magra y hambrienta;
> Piensa demasiado; los hombres así son peligrosos.

Las descripciones del científico Ernst Kretschmer (1925) son menos poéticas, pero quizá más sistemáticas que los escritos de autores famosos. Kretschmer concluyó que tanto la consti-

tución alta y delgada (*asténica*) como la constitución corporal musculosa (*atlética*) se asocian con tendencias al alejamiento (personalidad esquizoide). Por otro lado, se dice que una constitución corporal baja y robusta (*pícnica*) se asocia con inestabilidad emocional (personalidad cicloide). Sheldon, Stevens y Tucker (1940) propusieron una tipología relacionada. Su sistema somatotipo de tres componentes clasificó los físicos humanos de acuerdo con el grado de obesidad (*endomorfia*), musculosidad (*mesomorfia*) y delgadez (*ectomorfia*) (figura 14.2). Se supone que las estructuras corporales que representan los extremos de esos tres componentes están relacionadas con los siguientes tipos de temperamento: endomorfia con viscerotonía (sociable, alegre, ama la comida); la mesomorfia con la somatotonía (atlético, agresivo), y la ectomorfia con la cerebrotonía (introvertido, estudioso).

Las teorías de la constitución corporal son fascinantes, pero su validez científica es cuestionable. Existen muchas excepciones a las relaciones hipotetizadas entre el físico y el temperamento, y se han propuesto interpretaciones diferentes para esas relaciones. Además, los psicólogos contemporáneos objetan las tipologías porque colocan a las personas en categorías y les asignan etiquetas. La etiquetación no sólo enfatiza en exceso la causación interna de la conducta, sino que puede actuar como profecía que se cumple por sí misma en la cual la gente se convierte en lo que dice la etiqueta. De este modo, un individuo al que se etiqueta como *introvertido* puede ser abandonado por posibles amigos, ocasionando que se vuelva más aislado. De manera simi-

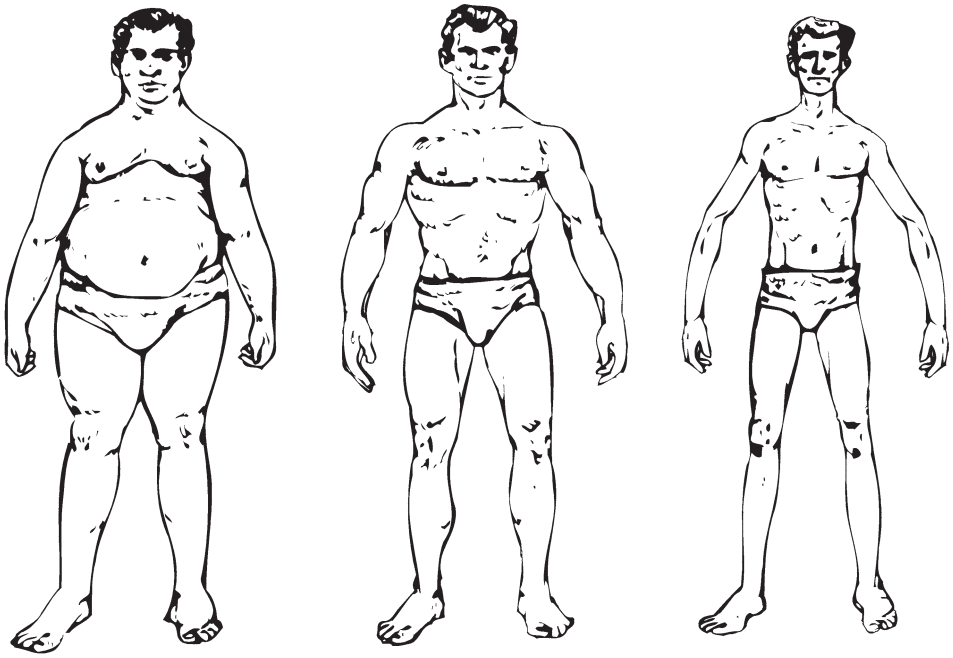


FIGURA 14.2 Somatotipos de Sheldon.

(Tomado de *Elements of Psychology*, de David Krech y Richard S. Crutchfield. Derechos reservados © 1958 por David Krech y Richard S. Crutchfield. Derechos reservados © 1969, 1974 por Alfred A. Knopf Inc. Reproducido con autorización.)

lar, un *extrovertido* puede volverse más comunicativo y sociable porque otra gente espera que se comporte de esa forma.

Teorías de los rasgos

Los rasgos de personalidad son menos generales que los tipos de personalidad. Gordon Allport, uno de los primeros teóricos de la personalidad, empezó su investigación sobre los rasgos al elaborar una lista de 17,953 palabras que en inglés se refieren a características de la personalidad, y al reducirla luego a una lista más pequeña de nombres de rasgos (Allport y Odbert, 1936). Allport definió el término *rasgo* como una "estructura neuropsíquica que tiene la capacidad de traducir muchos estímulos funcionalmente equivalentes, y de iniciar y guiar formas equivalentes (significativamente consistentes) de conducta adaptativa y expresiva" (Allport, 1961, p. 347). Para Allport, la *personalidad* consistía en la organización dinámica de esos rasgos que determinan el ajuste único de una persona al ambiente.

Otro teórico de los rasgos, R. B. Cattell, clasificó los rasgos en cuatro formas: comunes contra únicos, superficiales contra fuente, constitucionales contra moldeados por el ambiente, y dinámicos contra habilidad contra temperamento. Los rasgos comunes son características de toda la gente, mientras que los únicos son peculiares al individuo. Los rasgos superficiales de una persona pueden observarse con facilidad por su conducta, pero sus rasgos fuente sólo pueden ser descubiertos mediante procedimientos estadísticos de análisis factorial (vea el apéndice A). Los rasgos constitucionales dependen de la herencia, y los moldeados por el ambiente dependen del entorno. Por último, los rasgos dinámicos motivan a la persona hacia una meta, los rasgos de habilidad determinan la capacidad para alcanzar la meta, y los rasgos de temperamento atañen a los aspectos emocionales de la actividad dirigida hacia una meta. La teoría de los rasgos de Cattell, la cual es mucho más compleja de lo que sugiere esta breve descripción, ha servido como marco de trabajo para varios inventarios de personalidad, uno de los cuales es el Cuestionario de 16 Factores de la Personalidad.

Muchos otros psicólogos, incluyendo a Henry Murray, J. P. Guilford y Hans Eysenck, elaboraron teorías y realizaron investigación sobre los rasgos de personalidad. Los métodos del análisis factorial han sido aplicados a buena parte de esta investigación, arrojando una variedad de dimensiones de personalidad. Las dos dimensiones básicas del sistema de Eysenck, introversión-extroversión y estabilidad-inestabilidad se ilustran en la figura 14.3. Las posiciones de las 32 variables de personalidad en los ejes vertical y horizontal de esta figura indican la dirección y magnitud de esas características en las dos dimensiones.

Teoría psicoanalítica

Sigmund Freud y otros psicoanalistas veían a la personalidad humana como una especie de campo de batalla donde tres combatientes, el ello, el yo y el superyó, contienden por la supremacía. El *ello*, una reserva de pulsiones instintivas de sexo y agresión, albergado en la parte inconsciente de la mente, actúa de acuerdo con el principio del placer. Entra en conflicto con el *superyó* (la conciencia), que actúa de acuerdo con el principio moral. El ello es innato, pero el superyó se desarrolla a partir de la interiorización de las prohibiciones y sanciones establecidas por los padres sobre la conducta del niño. Mientras tanto, el *yo*, que funciona de acuerdo con el principio de la realidad, actúa como mediador entre las presiones implacables del ello y el superyó por el control. El ello dice "¡Ahora!", el superyó dice "¡Nunca!", y el yo dice "¡Más tarde!" a los deseos básicos del individuo. Los impulsos y el conflicto del ello con el superyó y el yo tienen lugar a menudo en la mente inconsciente, pero se expresan en pensamientos y conductas disfrazadas de varias formas.

FIGURA 14.3 Dimensiones de la personalidad, Eysenck.

(Tomado de *Personality and Individual Differences* de H. J. Eysenck y M. W. Eysenck, Plenum Publishing, Nueva York, 1958. Reproducido con autorización del editor.)

Freud también creía que la personalidad humana se desarrolla a través de una serie de *etapas psicosexuales*. Durante cada etapa, una región diferente del cuerpo (*zona erógena*) es el centro de estimulación y gratificación sexual, y en esa etapa predominan los conflictos que corresponden a la región corporal particular. La *etapa oral* ocurre desde el nacimiento hasta el año y medio de edad; en esta etapa el placer se deriva sobre todo de la estimulación de la boca y los labios, como al succionar, morder y tragar. Durante la *etapa anal*, desde alrededor del año y medio hasta los tres años, el interés y el conflicto se centran en la retención y expulsión de las heces. El negativismo, manifestado por el desafío a las órdenes de los padres y asociado con frecuencia al entrenamiento de control de esfínteres, es más pronunciado durante la etapa anal. En este orden sigue la *etapa fálica*, de los tres a los seis años, cuando la región corporal de mayor interés es el área genital. Durante esta etapa —cuando se enfatiza el frotarse, tocarse y exhibirse— se desarrolla el complejo de Edipo. El *complejo de Edipo*, considerado por Freud como un fenómeno universal, consta de un compuesto de sentimientos sexuales hacia la madre y disgusto por el padre en niños varones de tres a seis años de edad. La situación comparable en las niñas, disgusto por la madre y amor por el padre, se conoce como *complejo de Electra*.

Freud sostenía que para que un niño prosiga el desarrollo psicosexual después de la etapa fálica al *periodo de latencia* de relativa inactividad sexual durante la niñez media, el complejo de Edipo debe ser resuelto o reprimido. En la mayoría de los casos es resuelto cuando el niño aprende a identificarse con el padre, es decir, a tratar de actuar como él. Al inicio de la pubertad el niño que ha pasado con éxito por las etapas psicosexuales previas entra en la *etapa genital*. El interés por el sexo opuesto se vuelve dominante y, por lo general, culmina en el apareamiento heterosexual.

Freud fue uno de los primeros teóricos de la personalidad en reconocer que "el niño es el padre del hombre", que la privación y el conflicto en la niñez pueden tener efectos persistentes en la personalidad. Su teoría de las etapas psicosexuales sostiene que la frustración y el conflicto en una etapa particular afectan la estructura del carácter adulto ocasionando *fijación* (fracaso para progresar psicosexualmente más allá de una etapa particular) o *regresión* (regreso parcial o completo a un patrón de conducta típico de una etapa anterior del desarrollo). Por ejemplo, se dice que una persona que queda fijada en la etapa oral se caracteriza por la dependencia excesiva, la gula y la pasividad; una persona que está fijada en la etapa anal es excesivamente ordenada, obstinada y avara.

La teoría de Freud sobre la personalidad se basó casi por completo en observaciones clínicas no controladas de alrededor de 100 pacientes, y muchos rasgos de la teoría no han sido confirmados por la investigación. Ciertas suposiciones, por ejemplo, que el complejo de Edipo es universal y que existe un periodo de latencia en el desarrollo psicosexual, son incorrectas casi con certeza. De cualquier manera, sin duda Freud y sus seguidores estuvieron en lo correcto al enfatizar la existencia de la sexualidad infantil y su importancia para el desarrollo de la personalidad, el papel importante desempeñado por la motivación inconsciente en el moldeamiento de la personalidad y la conducta, y las funciones de los mecanismos de defensa al ayudar al individuo a afrontar la ansiedad producida por el conflicto intrapsíquico. Sin embargo, la idea de que los niños pasan de manera invariable por la secuencia de etapas psicosexuales bosquejadas arriba, y que la personalidad del adulto es moldeada por los conflictos sexuales de la niñez, es debatible y fue modificada por los psicoanalistas posteriores. En comparación con Freud, la teoría psicoanalítica moderna pone más énfasis en el aprendizaje social y la cultura que en los instintos biológicos como determinantes de la personalidad.

Teorías fenomenológicas

Producto de una tradición filosófica que enfatiza el análisis de la experiencia inmediata, personal y subjetiva, los teóricos fenomenológicos (humanistas o "del yo") sostienen que los teóricos de los rasgos y otros que intentan dividir la personalidad en un conjunto de componentes cometen una injusticia con su organización integrada y dinámica. En consecuencia, los teóricos fenomenológicos han sido críticos de los enfoques psicoanalítico, de rasgos y factores, y conductual para la comprensión de la personalidad. En contraste con el psicoanálisis tradicional, que enfatiza la importancia fundamental de los impulsos sexual y agresivo, el inconsciente y las etapas psicosexuales del desarrollo, los teóricos fenomenológicos subrayan las percepciones, los significados, los sentimientos y el yo. Ellos consideran que la gente responde al mundo en términos de sus percepciones únicas y privadas del mismo. Esas percepciones están determinadas por experiencias y los significados atribuidos a esas experiencias en un esfuerzo por realizar de manera plena el propio potencial. La parte del ambiente que se percibe y tiene significado para el individuo se conoce como *campo fenomenal*, una parte del cual (el *yo*) está relacionado con el individuo de una manera personal. Por último, la totalidad de las buenas y malas evaluaciones dadas por una persona al yo se conocen como *autoconcepto*.

De acuerdo con Abraham Maslow, Carl Rogers y otros teóricos fenomenológicos, todos pasamos por un proceso en el que nos esforzamos por alcanzar una congruencia o armonía entre el yo verdadero y el ideal, o *autorrealización.* La dirección básica de la existencia es hacia la autorrealización y las relaciones placenteras con los demás, pero este esfuerzo puede ser inhibido de varias maneras. Carl Rogers señalaba que la mayoría de las personas no están abiertas o dispuestas a aceptar toda la gama de sus experiencias. En el proceso de crecimiento aprenden que son objeto de *consideración positiva condicional*, en la cual su conducta es considerada aceptable por los padres y otras personas significativas sólo si se ajusta a los estándares aceptados (*condiciones de valor*). En consecuencia, el niño, que a la postre se convierte en adulto, aprende a reconocer y aceptar sólo una parte de sus experiencias. El resultado es un individuo que funciona de manera incompleta, y no puede funcionar de manera plena hasta que otras personas lo traten con *consideración positiva incondicional*. Esto es, cuando el individuo es aceptado independientemente de lo que hace.

Los clínicos que apoyan la teoría fenomenológica tienden a preferir los estudios de caso y las entrevistas no estructuradas en lugar de pruebas y procedimientos psicológicos objetivos. Carl Rogers no creía demasiado en el valor de los instrumentos de evaluación de la personalidad, y la teoría fenomenológica, o del yo, no ha sido tan influyente como las teorías de rasgos y factores y la psicoanalítica en el desarrollo de tales instrumentos. Aún así, muchos instrumentos y procedimientos para evaluar los sentimientos y actitudes hacia el yo se han basado en un concepto fenomenológico de la personalidad. Algunos ejemplos son las clasificaciones Q (Stephenson, 1953) y los inventarios como la Escala de Autoconcepto de Tennessee, la Escala de Autoconcepto para Niños de Piers-Harris y los Inventarios de Autoestima de Coopersmith.

Teoría del aprendizaje social

Muchos otros conceptos teóricos han influido en el desarrollo de los instrumentos de evaluación de la personalidad. Entre éstos se encuentra la teoría de George Kelly (1955) sobre los constructos personales y el enfoque cognitivo-conductual de los teóricos del aprendizaje social como Julian Rotter, Albert Bandura y Walter Mischel.

La teoría de Rotter. La primera *teoría del aprendizaje social* expuesta como tal fue la de Julian Rotter (1954), quien intentó integrar la posición conductista tradicional sobre el papel del reforzamiento en el aprendizaje con las conceptualizaciones cognoscitivas de Kurt Lewin y otros teóricos del campo. Rotter no fue el primero en advertir que la mayor parte de la conducta humana se aprende en un contexto social, pero hizo un esfuerzo más consciente que sus predecesores por desarrollar una teoría sistemática sobre la forma en que se lleva a cabo este proceso. Rotter distinguió entre reforzamientos y cogniciones: los reforzamientos producen movimiento hacia o lejos de una meta, mientras que las cogniciones son estados internos como las expectativas y el valor del reforzamiento. El término *expectativa* se refiere a una estimación que hace la persona de la probabilidad subjetiva de que una conducta específica realizada en cierta situación conducirá al reforzamiento. Dos *expectativas generalizadas* medidas e investigadas por Rotter y otros son el locus de control y la confianza interpersonal. El *locus de control* se refiere a la dirección típica a partir de la cual la gente percibe que es controlada (interna, o del interior de uno mismo, contra externa, o por otra gente). La *confianza interpersonal* atañe al grado en que una persona cree que otra gente dice la verdad.

De acuerdo con Rotter, el reforzamiento es importante para el desempeño, pero no todos los reforzamientos son valorados igualmente por el individuo. Aun cuando las probabilidades de ocurrencia de diferentes reforzamientos sean iguales, ciertos objetos o acciones tendrán mayor

valor de reforzamiento que otros. Tanto el valor del reforzamiento como las expectativas son afectados por la relevancia psicológica o significado de la situación para el individuo, y deben ser entendidos para poder predecir cómo se comportará la persona en dicha situación.

Teoría del aprendizaje por observación de Bandura. La teoría del aprendizaje social de Albert Bandura (1977) es más importante para el desarrollo de las técnicas destinadas a la modificación de la conducta inadaptada que para influir en el diseño de instrumentos de evaluación de la personalidad. Al conceptualizar el funcionamiento psicológico como interacciones recíprocas de variables conductuales, variables de la persona (cogniciones y otros estados internos) y variables ambientales, Bandura enfatiza que un ser humano no es un autómata pasivo ("push button") que sólo actúa cuando se actúa sobre él. Las personas influyen y son influidas por el ambiente social, en el cual tiene lugar el aprendizaje por medio de la observación, la imitación y el modelado. A diferencia de conductistas más tradicionales, como Clark Hull y B. F. Skinner, Bandura sostuvo que buena parte del aprendizaje tiene lugar sin reforzamiento, en ausencia de recompensas y castigos, pero que el reforzamiento es importante para determinar cuándo ocurre la conducta aprendida. De particular relevancia en el proceso de aprendizaje es el modelado de la conducta de otros. La efectividad del modelado depende de las características personales del modelo y del nivel de motivación del aprendiz. De acuerdo con Bandura, la agresión, los temores, las conductas de tipo sexual y muchas otras reacciones emocionales y de estilo se aprenden mediante la observación y el modelado.

Bandura también enfatizó el hecho de que el aprendizaje y la conducta son mediados por las percepciones y las cogniciones: la gente usa representaciones simbólicas, internas, de su ambiente, y estas representaciones median los cambios en la conducta. Al visualizar las consecuencias de sus acciones la gente aprende a regular su conducta.

Aproximaciones empíricas a la evaluación de la personalidad

En lugar de haber sido diseñados de acuerdo con una teoría formal de la personalidad, muchos instrumentos de evaluación de la personalidad han sido elaborados sobre una base puramente empírica. Por ejemplo, los reactivos de las diversas escalas del Inventario Multifásico de Personalidad de Minnesota (MMPI) se seleccionaron de acuerdo con su capacidad para distinguir entre dos grupos contrastantes de personas (normales y grupos de pacientes psiquiátricos seleccionados). En este procedimiento empírico no se involucró ninguna teoría específica de la personalidad; los reactivos del MMPI simplemente fueron validados contra el criterio específico de los diagnósticos psiquiátricos en varias muestras de pacientes mentales.

En el contexto de las investigaciones sobre la personalidad y los trastornos de conducta se han desarrollado muchos instrumentos de evaluación. Esos esfuerzos, aunque no carecen por completo de fundamentos teóricos, no han estado restringidos a una sola posición teórica. Algunos ejemplos de tales enfoques son los programas de investigación de los personólogos de Harvard y de los psicólogos del Instituto para la Evaluación e Investigación de la Personalidad en la Universidad de California en Berkeley.

USOS Y ABUSOS DE LA EVALUACIÓN DE LA PERSONALIDAD

Procedimientos e instrumentos de evaluación de la personalidad se utilizan en escuelas, clínicas, hospitales, prisiones y otros escenarios donde los resultados contribuyen a tomar decisiones

acerca de la gente. Idealmente, los resultados se tratan de manera cuidadosa y con plena conciencia de las limitaciones de las evaluaciones y de las necesidades y los derechos de los examinados. Por desgracia, la ética de los evaluadores de la personalidad no siempre es como debería ser.

Problemas éticos

Entre los métodos usados en la evaluación de la personalidad se encuentran las observaciones, las entrevistas, las escalas de calificación, las listas de verificación, los inventarios y las técnicas proyectivas. En ocasiones esos métodos han sido mal aplicados por personas no capacitadas o sin ética, lo que resulta en una marca negra para las pruebas psicológicas como un todo. No es difícil que una persona, habiendo leído un poco sobre psicología, obtenga unos cuantos instrumentos de lápiz y papel y pretenda ser un analista de la personalidad. Al igual que los adivinos y otros charlatanes, estos supuestos diagnosticadores de la personalidad manejan generalidades, trivialidades y otras afirmaciones que parecen específicas para un individuo, pero que en realidad se aplican a la mayoría de la gente. Para demostrar este "efecto Barnum", considere el siguiente "perfil de personalidad":

> Tiene una fuerte necesidad de agradar a otra gente y que ésta lo admire. Tiene la tendencia a ser crítico consigo mismo. Tiene una gran capacidad que no ha utilizado y que no ha aprovechado. Aunque tiene algunas debilidades de personalidad, por lo general puede compensarlas. Su ajuste sexual le ha causado algunos problemas. Aunque disciplinado y controlado en su exterior, en su interior tiende a preocuparse y a ser inseguro. En ocasiones tiene serias dudas respecto a si ha tomado la decisión correcta o si ha hecho lo adecuado. Prefiere cierta cantidad de cambio y variedad y se siente insatisfecho cuando se enfrenta a restricciones y limitaciones. Se enorgullece de ser un pensador independiente y no acepta las opiniones de otros sin prueba satisfactoria. Se ha percatado de que es poco prudente ser demasiado franco al revelarse ante los demás. En ocasiones es extrovertido, afable y sociable, mientras que otras veces es introvertido, cauteloso y reservado. Algunas de sus aspiraciones tienden a ser muy poco realistas (Forer, 1949).

¿Es ésta una buena descripción de su personalidad? 37 estudiantes en un grupo de 50 a quienes presenté el párrafo anterior lo calificaron como una descripción buena o excelente de su personalidad.

Se necesita mucho entrenamiento y experiencia para convertirse en un buen observador e intérprete de la personalidad humana. Los maestros, gerentes de personal y otras personas que no son psicólogos, a menudo pueden aplicar escalas de calificación y listas de verificación de una manera sensible, pero la aplicación e interpretación de inventarios de personalidad y técnicas proyectivas están restringidas a los psicólogos y otros profesionales con formación comparable. Incluso entonces es cuestionable la utilización de muchos inventarios de personalidad y técnicas proyectivas con propósitos que no sean de investigación. Muy a menudo, las pruebas de personalidad se concentran más en los síntomas de desadaptación y enfermedad mental que en los de adaptación y salud mental. Debido a que esos temas son muy personales y deben manejarse con cuidado, es necesario ser cauteloso al aplicar e interpretar los resultados de cualquier instrumento de evaluación de la personalidad. Se debe respetar tanto el derecho del examinado a la privacidad como su preocupación natural por su estabilidad emocional y salud mental.

Además de la cuestión de la privacidad y otros temas morales, la confiabilidad y validez de los instrumentos de evaluación de la personalidad plantean problemas. Responder a los materiales de prueba dando las respuestas que son socialmente más deseables o dejando que las res-

puestas sean determinadas por cualquier papel que se sienta necesario asumir en la situación específica de prueba son tipos de grupos de respuesta que pueden invalidar los resultados de una evaluación de la personalidad. El uso de medidas más discretas que resulten menos susceptibles a la simulación o en las cuales no interfiera indebidamente el procedimiento con la obtención de resultados válidos puede ser una solución para los grupos de respuesta, pero al mismo tiempo introduce los problemas de cuantificación e interpretación de los resultados.

Interpretación de los datos de la evaluación

Aun cuando se utilicen medidas de personalidad elaboradas y validadas con cuidado, es una regla fundamental que las interpretaciones resultantes deben verse como hipótesis a ser confirmadas o refutadas por información subsecuente. Los resultados de una evaluación de la personalidad no son exactos ni finales, y pueden verse de maneras diferentes por distintos examinadores. Esto se vuelve embarazosamente obvio cuando diferentes psicólogos o psiquiatras, que actúan como testigos expertos en un caso legal, difieren de manera radical en sus interpretaciones acerca de los mismos resultados de una evaluación. Dada la naturaleza subjetiva de la mayoría de las evaluaciones psicológicas, tales vergüenzas pueden ser inevitables cuando dos partes en una disputa legal tienen objetivos diferentes.

Pueden hacerse varias recomendaciones adicionales que conciernen a la obtención e interpretación de datos de evaluación de la personalidad (adaptados de Sundberg, 1977):

1. Estudie la situación general de la vida y los problemas del examinado, y luego obtenga más detalles en áreas de relevancia particular para la evaluación.
2. Sea sensible a los antecedentes socioculturales y étnicos del examinado, así como a su edad y sexo si es relevante.
3. Siempre que sea posible utilice las técnicas y los datos más objetivos, en lugar de los subjetivos.
4. Obtenga el tipo correcto de información, no sólo más información, acerca de la situación específica y los propósitos de la evaluación.
5. Evite formular demasiada especulación al interpretar los resultados y predecir la conducta; tenga especial cuidado al hacer predicciones concernientes a comportamientos con baja probabilidad de ocurrencia.
6. De ser posible, verifique sus hallazgos e interpretaciones con los de otros asesores psicológicos, y lleve un registro de sus acuerdos y desacuerdos, éxitos y fracasos.
7. Comunique los resultados por escrito en un estilo que pueda ser entendido por la gente a quien se dirige el informe.

Informe de los resultados de la evaluación

Cualesquiera que sean las razones para realizar un examen psicológico, por lo regular se requiere algún tipo de informe escrito de los resultados. La reseña y longitud del informe de un caso clínico de estudio varían de acuerdo con los propósitos del estudio y los lectores a quienes se dirige el informe, pero la forma 14.1 proporciona detalles acerca de los tipos de información incluidos en dicho informe.

Al preparar el informe formal de un examen psicológico, quien lo escribe debe tener claras las preguntas de canalización o las quejas principales (por qué fue canalizada la persona a evaluación psicológica o por qué la buscó) y cómo responden a esas preguntas o dan solución a esos

FORMA 14.1 Formato de un informe de evaluación psicológica

Nombre del examinado _____

Edad _____ Fecha de nacimiento _____ Educación _____

Examinador _____

Lugar del examen _____ Fecha _____

Pruebas y otros procedimientos usados. Mencione los nombres, incluyendo las formas y los niveles, de todas las pruebas e inventarios que haya aplicado.

Información sobre la canalización y antecedentes. ¿Por qué se canalizó al examinado a evaluación psicológica? ¿Cuál fue el propósito de la canalización y qué persona u organización la hizo? ¿Qué información antecedente relevante para el caso fue obtenida de otras fuentes (registros escolares, entrevistas y cuestiones similares)? Exponga la historia del examinado, así como la de otros observadores si está disponible. Describa la historia física y psicológica del examinado, así como sus características y situación educativa y de empleo. En el caso particular de los niños, es importante la información sobre el hogar y la familia (posición social, características de los padres, hermanos, etc.). También deben advertirse impedimentos sensoriales o psicomotrices serios, así como la presencia de trastorno emocional.

Apariencia y características conductuales. Describa la apariencia y la conducta del examinado durante el examen. Describa las características del examinado, su forma de abordar las tareas, su nivel de motivación y emocionalidad, y cualquier otro factor que pueda haber influido en los resultados del examen. ¿Qué conductas del examinado fueron sintomáticas de condiciones o características físicas, cognoscitivas o afectivas particulares?

Resultados e interpretaciones. Proporcione una descripción detallada de los resultados de las pruebas u otros instrumentos administrados y la forma en que pueden interpretarse. Si el examinador está interpretando los resultados de acuerdo con una teoría particular de la personalidad o la conducta, asegúrese de que el lector entiende el lenguaje y las suposiciones de la teoría. Sea lo más específico posible al interpretar los resultados.

Conclusiones y recomendaciones. Describa las conclusiones (descriptivas, dinámicas, de diagnóstico) que se deriven de los datos de las observaciones, entrevistas y pruebas estandarizadas o no estandarizadas. ¿Qué recomendaciones garantizan los resultados? Incluya las advertencias apropiadas para la interpretación, pero no caiga en generalidades. Entre las recomendaciones que pueden hacerse se encuentran la evaluación psicológica adicional (sea específico), exámenes neurológicos u otros exámenes médicos, asesoría o psicoterapia, colocación en clases o entrenamiento especiales, rehabilitación vocacional e institucionalización. En caso de existir un impedimento o discapacidad, ¿tiene remedio?

_____

Firma del examinador

problemas los resultados de la evaluación. También debe considerarse proporcionar información sobre la condición mental presente y la estabilidad emocional, así como los probables resultados (pronóstico) de la condición del paciente. Las características del examinado y sus interrelaciones deben describirse de manera tan completa y específica como sea posible, evitando generalizaciones vagas, estereotipos y banalidades. También es útil para la persona que hace

el informe tener una teoría de la personalidad, o al menos un marco de referencia psicológico, sobre la que pueda basar su interpretación de los resultados. Por último, el informe debe escribirse en un estilo conciso y claro que sea comprensible para el lector. Un informe psicológico es de poco valor si no lo entienden o lo leen quienes están en posición de usar la información para tomar decisiones que atañen a la vida y el bienestar del examinado.

EVALUACIÓN CLÍNICA

Aunque por lo general los psicólogos clínicos dedican más tiempo a actividades de tratamiento, consulta, investigación, enseñanza y otras similares que a la evaluación psicológica, muchos clínicos encuentran útiles pruebas objetivas como el MMPI y técnicas proyectivas como el Rorschach para elaborar el psicodiagnóstico y la planeación del tratamiento. La evaluación clínica con propósitos de identificación y diagnóstico de trastornos de la conducta y cognoscitivos, y para la planeación de tratamientos u otros procedimientos de intervención, tiene lugar en una variedad de escenarios. Éstos incluyen oficinas privadas, clínicas de salud mental, hospitales mentales, centros médicos de la Administración de Veteranos en Estados Unidos, escuelas, instituciones de custodia y escenarios forenses. Los psicólogos clínicos son llamados para realizar evaluaciones psicológicas en: escenarios de salud mental con propósitos de diagnóstico, tratamiento y ubicación residencial; escenarios médicos como auxiliares en la evaluación de los aspectos psicológicos de la enfermedad; escenarios de tratamiento como auxiliares en la planeación y evaluación de la efectividad de la psicoterapia y la quimioterapia; ambientes educativos como un apoyo en la formulación de medidas apropiadas de remedio; entornos legales para ayudar a las autoridades en audiencias sobre demencia, decisiones de custodia y planeación de medidas de rehabilitación, y en varios escenarios para realizar evaluaciones requeridas por la ley, como en los casos que involucran compensación estatal o federal.

Una vez reconocida la necesidad de evaluación clínica, pueden establecerse las metas y tomarse decisiones concernientes a los tipos de datos que se necesitan para alcanzarlas. Las metas generales de la evaluación clínica son proporcionar una descripción precisa de la problemática del paciente (cliente), determinar qué factores interpersonales y ambientales la precipitaron y están manteniendo, y efectuar predicciones acerca de los resultados con y sin intervención. La obtención del tipo de información requerida en escenarios clínicos a menudo demanda un estudio de caso minucioso.

Examen del estado mental y estudio de caso clínico

A los psicólogos clínicos se les pide con frecuencia que realicen un *examen del estado mental* para obtener información pormenorizada acerca del estado emocional de una persona (afecto y estado de ánimo), de su funcionamiento intelectual y perceptual (atención, concentración, memoria, inteligencia y juicio), del estilo y contenido de sus procesos de pensamiento y habla, del nivel de introspección sobre su estado mental y sus problemas de personalidad y actividad psicomotriz, así como de la apariencia general de la persona, su actitud e introspección acerca de su condición. No toda esta información se obtiene de pruebas psicológicas. También se requieren observaciones cuidadosas y entrevistas detalladas de la persona y de quienes la conocen bien. Al conducir un estudio minucioso de caso, el cliente y otras personas significativas proporcionan información sobre los antecedentes y el carácter, y se recaban datos de seguimiento a lo largo de

cierto periodo. Puede solicitarse información sobre la familia, la cultura, el historial médico, educativo y del desarrollo, la posición económica, los antecedentes legales y las actividades y pensamientos de la persona. Luego de obtener e integrar los datos de la evaluación, se prepara un informe que resume los hallazgos y describe las fortalezas y debilidades de la persona. En el informe pueden anotarse las recomendaciones pertinentes para efectuar intervenciones clínicas, educativas o vocacionales.

Cuando se conduce un estudio de caso para determinar la causa o las causas de un problema psicológico específico, pueden formularse hipótesis o conclusiones concernientes a la causación, y emitirse recomendaciones específicas que atañen al tratamiento (psicoterapia, medicamentos u otro tratamiento médico, educación especial, etc.). Después de un intervalo apropiado debe hacerse una evaluación de seguimiento para conocer la efectividad del programa de tratamiento prescrito.

A pesar de que arroja información potencialmente útil para formular una imagen global, y de que proporciona una comprensión profunda del individuo, un estudio de caso clínico tiene varias debilidades notables. Éstas incluyen la naturaleza introspectiva de los datos (la memoria rara vez es precisa), el hecho de que el conductor del estudio con frecuencia muestra sesgos al seleccionar y evaluar ciertos tipos de datos o mediciones, y la generalización limitada de los hallazgos entre situaciones o circunstancias encontradas por la persona. El empleo de una variedad de evaluaciones en una muestra sistemática de situaciones y estar consciente de la probabilidad de sesgo en la selección y evaluación puede ayudar a reducir, si no es que a eliminar, las malas interpretaciones y las generalizaciones excesivas.

Psicodiagnóstico

El *psicodiagnóstico* es un proceso mediante el cual se examina a una persona desde un punto de vista psicológico para determinar la naturaleza y el grado de un trastorno mental o conductual. En el *modelo médico* tradicional de los trastornos mentales, la persona que hace el psicodiagnóstico observa, entrevista y prueba al paciente para determinar la presencia o ausencia de ciertos síntomas psicológicos (y físicos). Quien hace el diagnóstico compara luego los síntomas del paciente con descripciones estándar de la conducta anormal para determinar a qué categoría de trastornos se ajusta mejor el paciente. El resultado final de este proceso es la asignación de una clasificación psiquiátrica al paciente, según se especifica en el *Manual Diagnóstico y Estadístico de Trastornos Mentales-IV* (*DSM-IV*) (American Psychiatric Association, 1994) o en la *Clasificación Internacional de Enfermedades* (*ICD-10*) (World Health Organization, 1992). Además de diagnosticar el trastorno, se hace un *pronóstico* o predicción del probable resultado.

La habilidad para emitir psicodiagnósticos precisos requiere de gran entrenamiento y experiencia, e incluso entonces puede ser considerable la probabilidad de cometer un error. Los clínicos cometen errores en el diagnóstico debido a la percepción selectiva, el recuerdo selectivo, experiencia insuficiente, seguimiento inadecuado y lógica deficiente.

Arkes (1994) describe una serie de errores cometidos en el psicodiagnóstico y otros juicios clínicos. Una fuente de error es el *problema de la correlación ilusoria*, que consiste en basar los juicios clínicos en el número de veces que cierto signo o indicador y un trastorno específico han ocurrido juntos, pero pasar por alto el hecho de que no han ocurrido juntos incluso con mayor frecuencia. Los encargados de elaborar los psicodiagnósticos cometen este error cuando advierten o recuerdan cualquier cosa que se ajuste a sus expectativas, pero ignoran u olvidan lo que sea contrario a dichas expectativas.

Una segunda fuente de error en los juicios clínicos puede ocurrir cuando se desconoce la *tasa base*, la proporción de personas en una población particular que posee una característica o condición específica. Debido a la operación del azar, es mucho más fácil identificar un signo diagnóstico particular o predecir cierto tipo de conducta cuando su tasa base es alta que cuando es baja. Por ejemplo, la tasa de suicidios en la población general es relativamente baja, pero la tasa de conducta neurótica es bastante alta. En consecuencia, es más fácil predecir la conducta neurótica que el suicidio.

Una tercera fuente de error que le resta méritos a los juicios clínicos es el *sesgo retrospectivo* de creer que después de ocurrido un evento, alguien podría haberlo anticipado si se le hubiera pedido hacerlo. Un ejemplo de este error es concluir que un conocido que ha cometido un acto violento siempre fue abiertamente agresivo y perturbado.

Una cuarta fuente de error en el juicio clínico es la *confianza excesiva* en los juicios propios a pesar de evidencia que los contradiga. Por ejemplo, la simple formulación de una regla como "los psicóticos son pálidos" puede ser suficiente para convencer al clínico de que tal síntoma es válido.

Conferencia de caso

Un informe por escrito es sólo una forma en que los resultados de una evaluación psicométrica se comunican a quienes tienen el derecho legítimo de conocerlos. Las conferencias de casos clínicos o consultas en contextos de salud mental, y conferencias entre padres y maestro o padres y consejero en escenarios escolares, pueden ocurrir antes y después de una evaluación psicológica. Cuando se conduce una conferencia posterior a la prueba con una persona que desconoce la terminología psicológica, como un padre típico, el examinador debe describir, en un lenguaje apropiado para el escucha, los resultados de la prueba y las conclusiones que puedan derivarse razonablemente de ellos. En general, deben emplearse descripciones e interpretaciones cualitativas más que cuantitativas. También deben analizarse el propósito y la naturaleza de las pruebas, el motivo de su selección en particular, y las limitaciones de las pruebas y de sus resultados. Deben usarse declaraciones descriptivas, más que etiquetas, y rangos de calificación que consideren el error estándar de medición más que calificaciones específicas. La consulta incluye también un análisis de las opciones y decisiones de tratamiento, remedio, rehabilitación u otra intervención y el proporcionar información sobre las fuentes de canalización. Después de la consulta, el examinador debe enviar una copia del informe del examen a la fuente de canalización y a otras partes responsables que tengan necesidad y derecho a conocerlo.

OTRAS ÁREAS DE APLICACIÓN DE LA EVALUACIÓN DE LA PERSONALIDAD

Durante la década pasada se observó una mayor demanda de servicios psicológicos en otras tres áreas, siendo éstas el matrimonio y la familia, la salud y los asuntos legales. Dichos ámbitos han atraído la atención de los psicólogos investigadores y de otros profesionales interesados en desarrollar instrumentos psicométricos para investigación y aplicaciones en esas áreas. Muchos colegios y universidades ya han establecido incluso los programas de posgrado pertinentes, y la medición e investigación que les atañe son extensas.

Evaluación matrimonial y familiar

Varios instrumentos psicométricos han demostrado ser útiles en la identificación, el diagnóstico y el pronóstico concernientes a los problemas matrimoniales y familiares. Se dispone de listas de verificación, escalas de calificación, inventarios y otros instrumentos para el asesoramiento prematrimonial, la identificación de fuentes y posibles soluciones a desacuerdos y problemas familiares, y para ayudar a las víctimas del divorcio, padres e hijos, a recuperarse y seguir con sus vidas. A menudo se aplican inventarios y técnicas proyectivas tradicionales como el MMPI, el Cuestionario de 16 Factores de la Personalidad (16 FP), la prueba de manchas de tinta de Rorschach y la Prueba de Apercepción Temática (TAT), para analizar problemas matrimoniales y familiares. También se dispone de listas de verificación (por ejemplo, la Lista de Verificación de Evaluación Conyugal), inventarios (por ejemplo, la Evaluación de la Actitud Matrimonial, el Inventario de Satisfacción Conyugal y la Medida de Evaluación Familiar) y técnicas proyectivas especiales (por ejemplo, el Test de Relaciones Familiares: Versión para Niños y el Test de Apercepción Familiar). Otro instrumento psicométrico útil, la Escala del Entorno Familiar (Consulting Psychologists Press), se diseñó para evaluar el clima social de los sistemas familiares y determinar cómo interactúan las características de la familia. Este cuestionario puede utilizarse para identificar fortalezas, problemas y otros aspectos importantes en el tratamiento de la familia. La información proporcionada por todos los instrumentos anteriores debe ser complementada con entrevistas y observaciones sensibles de las parejas y los miembros de la familia en interacciones sociales cara a cara.

Psicología de la salud

La *psicología de la salud* ha sido definida como las "contribuciones educativas, científicas y profesionales de la disciplina de la psicología a la promoción y el mantenimiento de la salud, la prevención y el tratamiento de la enfermedad, y la identificación de los correlatos etiológicos y diagnósticos de la salud, la enfermedad y la disfunción relacionada" (Matarazzo, 1980, p. 815). El interés en el papel de las actitudes, la autoeficacia y otros factores psicológicos o variables de personalidad en la salud no se limita a los trastornos psicosomáticos, como las úlceras duodenales y las migrañas, sino que incluye los trastornos cardiovasculares, el cáncer y otras enfermedades que amenazan la vida. Los psicólogos son llamados no sólo para identificar factores psicológicos relacionados con varias condiciones médicas y para ayudar a diagnosticar trastornos específicos, sino también para auxiliar en la planeación de tratamientos o de otros procedimientos de intervención. El campo de la *medicina del comportamiento*, una subespecialidad de la psicología de la salud, ha hecho contribuciones significativas al tratamiento y manejo de pacientes con técnicas de modificación de conducta y otros procedimientos. El concepto de salud también ha alcanzado una connotación más amplia que la simple ausencia de enfermedad. Tal como es usada por los científicos sociales y conductuales, en particular, la salud se refiere ahora al *bienestar positivo* y a la obtención de una buena *calidad de vida.*

Se dispone de varios inventarios de personalidad relacionados con la salud para ayudar en la formulación de planes comprensivos de tratamiento para pacientes médicos adultos. Entre éstos se encuentran el Inventario de Uso del Alcohol (NCS Assessments) para evaluar la naturaleza del patrón de uso de alcohol de un individuo, el Inventario de Trastornos Alimenticios-2 (Psychological Assessment Resources) para evaluar rasgos conductuales asociados con la anorexia y la bulimia, y el Inventario de Salud Conductual de Millon (NCS Assessments) para ayudar en la formulación de planes comprensivos de tratamiento para pacientes médicos adultos. En

los años recientes se ha incrementado considerablemente el número de listas de verificación, escalas de calificación y otros cuestionarios sobre asuntos relacionados con la salud de los que se dispone de manera comercial. Entre éstos se incluyen instrumentos diseñados para identificar problemas de salud en general y áreas de salud específicas, cuestionarios de opiniones y creencias que atañen a la salud, medidas del estrés y formas de afrontarlo, medidas de percepción y control del dolor, y medidas de ansiedad, depresión, abuso de sustancias, violencia y suicidio potencial.

Las *atribuciones*, o explicaciones que la gente proporciona para las causas (internas o externas) de su conducta, están relacionadas con la eficacia y el control personales. Dos instrumentos diseñados para estudiar el papel de las atribuciones, y el concepto relacionado de *locus de control*, en la determinación de la conducta son la Prueba de Atribución de la Salud (de IPAT) y la Escala de Locus de Control de la Salud (Wallston y Wallston, 1981).

Otra variable que ha jugado un papel central en el campo de la psicología de la salud es el *estrés*. Entre los instrumentos comercialmente disponibles con la palabra *estrés* en el título están el Inventario del Estrés Cotidiano y el Índice de Estrés de los Padres (ambos de Psychological Assessment Resources). Otros instrumentos relacionados con el estrés son el Inventario de Recursos de Afrontamiento (Consulting Psychologists Press) y la Escala de confusión y exaltación (Mind Garden). Relacionado con la medición del estrés o con las reacciones al estrés está el campo de la *toxicología del comportamiento*, el cual se ocupa de la evaluación del desempeño bajo circunstancias ambientales adversas.

Psicología legal

La *psicología legal* se interesa en los aspectos psicológicos de la ley y en su cumplimiento. Los psicólogos que son empleados en contextos de cumplimiento de la ley, por lo regular son clínicos que poseen una amplia gama de habilidades y realizan gran variedad de tareas. Pueden usar pruebas, cuestionarios y procedimientos de entrevista para ayudar a seleccionar al personal que se encarga de hacer cumplir la ley. Pueden servir como expertos en relaciones humanas y formadores de equipos, conducir talleres y entrenar a oficiales de policía en técnicas de intervención en crisis como peleas domésticas y toma de rehenes. Pueden dar consejo o conducir psicoterapia de grupo e individual con los oficiales y sus familias. También pueden contribuir a la evaluación del desarrollo de programas en contextos de cumplimiento de la ley y realizar investigación sobre el entrenamiento y tratamiento del personal encargado del cumplimiento de la ley.

Una rama de la psicología legal, conocida como *psicología forense*, se interesa sobre todo en la evaluación de acusados en juicios legales para determinar si son competentes para enfrentar un juicio, si son peligrosos y/o si es probable que sean reincidentes. En los juicios legales tanto el acusador como la defensa pueden pedir a los psicólogos que examinen al demandante en búsqueda de signos de trastorno mental, conducta peligrosa o violenta, incompetencia para enfrentar el juicio o manejar sus propios asuntos, incapacidad para servir como padre adecuado en una audiencia de custodia infantil o para muchos otros propósitos.

No sólo el acusado sino también otras personas (testigos y otros) asociadas con una disputa legal pueden requerir un examen psicológico. Puede pedirse la opinión de los psicólogos concerniente a un delincuente desconocido o arrestado, si un niño estará mejor si se le coloca con uno de los padres o con otra persona, e incluso cómo es probable que voten los jurados potenciales en un juicio específico. Por ejemplo, puede pedirse a un experto en análisis de la personalidad que colabore en el proceso de selección de jurados en juicios penales o civiles.

Competencia y demencia. En los años recientes se han buscado cada vez con más frecuencia las opiniones y recomendaciones de los psicólogos en asuntos que tienen que ver con el tema de la competencia (competencia para enfrentar un juicio, compromiso civil, comprensión de los derechos Miranda y temas relacionados). La competencia para enfrentar un juicio tiene que ver con si un acusado entiende los cargos en su contra y si puede ayudar en su propia defensa.[1] Como afirmó la Suprema Corte de Estados Unidos en el caso Dusky contra Estados Unidos (1960), el acusado debe poseer "la capacidad suficiente para consultar con sus abogados con un grado razonable de comprensión racional... [y] de los hechos de los procedimientos en su contra". Esto significa que por lo general, pero no siempre, las personas que son retrasadas mentales, psicóticas o que sufren de algún trastorno neurológico debilitante, son consideradas incompetentes para enfrentar un juicio. Sin embargo, la incompetencia no es sinónimo de *demencia*. Mientras que la demencia legal atañe al estado mental del acusado en el momento que se cometió el delito, la condición de incompetencia es continua. Una persona puede ser encontrada "competente para ser sometida a juicio", pero puede ser declarada "no responsable por razones de demencia".

La Regla M'Naughten, la decisión Durham y el Código Penal Modelo han influido en las pruebas legales de demencia en Estados Unidos. Aunque los alegatos de demencia son admitidos en la mayoría de las entidades federativas de Estados Unidos, algunas los han abolido por completo. El estándar de demencia legal aplicado con mayor frecuencia por el sistema legal en Estados Unidos es el Código Penal Modelo propuesto por el Instituto Legal de Estados Unidos (ALI) y adoptado en 1972. La definición del ALI afirma:

> Una persona no es responsable de la conducta criminal, es decir, es demente, si en el momento de dicha conducta, como resultado de enfermedad o defecto mental, carece de capacidad sustancial para apreciar la criminalidad (ilegalidad) de su conducta o para ajustar su conducta al requerimiento de la ley (American Law Institute, 1956).

Entre los procedimientos y herramientas utilizados por los psicólogos para evaluar la competencia se encuentran guías de entrevista e instrumentos de detección de competencia como la Entrevista de Observación de Georgetown de la Competencia para Enfrentar un Juicio (Bukatman, Foy y De Grazia, 1971), la Prueba de Observación de Competencia (Lipsitt, Lelos y McGarry, 1971), el Instrumento de Evaluación de la Competencia (McGarry *et al*., 1973) y la Prueba de Competencia de la Corte de Georgia (Wildman *et al*., 1980). Las Escalas Rogers de Evaluación de la Responsabilidad Criminal (Psychological Assessment Resources) pueden aplicarse para determinar la responsabilidad criminal de acuerdo con el grado de deterioro psicológico que es significativo para declarar la demencia bajo el estándar del ALI. Las cinco escalas de este instrumento evalúan la confiabilidad del paciente, su organicidad, psicopatología, control cognoscitivo y control conductual en el momento que se supone el paciente cometió el delito. También pueden aplicarse pruebas neuropsicológicas a los acusados en alegatos por demencia.[2]

Dos de las pruebas aplicadas de manera más común en los contextos forenses son el MMPI y el Rorschach (vea los capítulos 17 y 18). Además de sus muchas otras aplicaciones en juris-

[1]Una clase de competencia parcial, la *capacidad testamentaria*, se refiere específicamente a la competencia para hacer un testamento. Un individuo que posee capacidad testamentaria, la cual se determina de manera legal, conoce la naturaleza y extensión de su propiedad, sabe que está haciendo un testamento y quiénes son sus beneficiarios naturales.

[2]Consulte a Rogers y Shuman (2000) para obtener información más detallada sobre la realización de evaluaciones de demencia.

prudencia, el MMPI puede contribuir a la identificación de la posición defensiva (poca disposición a decir la verdad) y proporcionar información concerniente a asuntos adicionales de la conducta personal que son de interés para los juicios legales. El Rorschach es otro intrumento valioso en los escenarios legales, pero ni éste ni el MMPI permiten respuestas y opiniones no calificadas concernientes a los asuntos legales.

Sexo y violencia. En lo que respecta a delitos sexuales, el Cuestionario Clarke sobre Antecedentes Sexuales para Varones (Langevin, 1983), diseñado para evaluar el tipo y la fuerza de la conducta sexual anómala, puede ser de ayuda para los psicólogos forenses.

Aunque no se ha desarrollado una prueba que por sí misma pueda predecir la conducta violenta, el MMPI puede contribuir a pronosticar un comportamiento peligroso o violento. También es útil la Lista de Verificación de Psicopatía de Hare, revisada (R. D. Hare; Psychological Assessment Resources), la cual es de gran aplicación. Una serie de indicadores conductuales, como una historia reciente de violencia, abuso de sustancias, ruptura de un matrimonio o una relación amorosa, disciplina o terminación en el trabajo y acceso a armas de fuego, también pueden contribuir a la predicción de la conducta violenta (Hall, 1987). Puede utilizarse una combinación de la historia personal y datos de pruebas para hacer una estimación de la probabilidad de ocurrencia de un comportamiento violento. La determinación del potencial para la conducta violenta es importante no sólo en las audiencias de libertad condicional y otros asuntos concernientes a los delincuentes convictos, sino también en la selección y promoción de oficiales de policía y otros guardianes de la paz.

La violencia puede ser expresada hacia adultos o niños, pero en años recientes el sistema legal y la sociedad como un todo se han sensibilizado a los alegatos de abuso físico contra los niños. En estos casos las observaciones, entrevistas, pruebas de dibujo de figuras y juego con muñecos pueden contribuir a la determinación o predicción de abuso físico o sexual de los niños.

Custodia de los hijos. Las evaluaciones para otorgar la custodia de los hijos pueden implicar entrevistas con los padres que se concentren en las prácticas de crianza infantil, además de la aplicación de pruebas de inteligencia y personalidad. Las mediciones del conocimiento y de las actitudes de los padres concernientes a las prácticas de crianza infantil también pueden contribuir a tomar decisiones en los casos de custodia de los hijos. El Cociente de Custodia, Gordon y Peck (1989), el cual arroja calificaciones en 10 factores de paternidad, puede ser útil a este respecto.

Un sistema inclusivo para la evaluación de padres e hijos en casos de abuso o maltrato de los niños es el Sistema de Evaluación Uniforme para la Custodia de los Hijos (Psychological Assessment Resources). Una evaluación completa con este sistema supone completar diez formularios de datos generales y administrativos, nueve formularios para los padres y seis para el niño. En los formularios para los padres se incluyen una historia familiar-personal completa, dos entrevistas, una lista de verificación de las habilidades del padre, observaciones de las interacciones padre-hijo, una visita de observación al hogar y algunos otros formularios.

La evaluación de los niños en los casos de custodia puede implicar la aplicación de instrumentos psicométricos estandarizados como la subprueba Comprensión de la WPPSI-R o WISC-III, pruebas de relato de cuentos y las Escalas Perceptuales de Bricklin (Bricklin, 1984). El último instrumento se concentra en entender las percepciones que tiene el niño de sus padres en cuatro áreas: competencia, apoyo, consistencia de seguimiento y posesión de rasgos de personalidad admirables. Se acostumbra hablar con los niños y quizá emplear otras técnicas (juego

con muñecos y dibujo de figuras concernientes a las situaciones de la vida familiar, pruebas de frases incompletas, etc.) para determinar si tienen alguna preferencia con respecto a los arreglos de residencia y visita futuros. Sin embargo, debe reconocerse que las preferencias e informes manifestados por preescolares de inteligencia promedio o inferior al promedio con frecuencia no son muy confiables, y están demasiado influidos por acontecimientos recientes a los que los niños dan mucho valor.

PROBLEMAS Y CONTROVERSIAS EN LA EVALUACIÓN DE LA PERSONALIDAD

Al igual que las medidas de las habilidades cognoscitivas, la medición de la personalidad ha recibido críticas de psicólogos y no psicólogos. Debido quizá a que sus aplicaciones son menos extensivas y menos cruciales, los instrumentos de evaluación de la personalidad no han sido tan criticados como las pruebas de habilidad por el público general. Sin embargo, las características de medición relativamente pobres de muchas pruebas de personalidad no han dejado de ser advertidas por los profesionales y el público. Entre los no psicólogos que han denunciado las pruebas de personalidad están ciertos escritores y padres que objetan cuestiones o aproximaciones particulares usadas en la evaluación de características personales, actitudes y conducta.

Quema de pruebas en Texas

Algunos de los comentarios negativos más extremos en relación a esos instrumentos se encuentran en libros de Whyte (1956) y Gross (1962, 1965) sobre las aplicaciones de las pruebas de personalidad en los negocios y la industria. Por supuesto, White y Gross no fueron los primeros en encontrar defectos en los instrumentos de evaluación de la personalidad. Un indicador de los sentimientos por parte del público lego fue la quema de ciertas escalas de actitudes y otros cuestionarios y pruebas por orden del Consejo Escolar de Houston en 1959. La hoguera fue consecuencia de la enérgica protesta de un grupo de padres de Houston que objetaron el hecho de que, como parte de una investigación, se pidiera a sus hijos que respondieran "cierto" o "falso" a reactivos como:

> Me gusta sumergirme en la bañera.
>
> Cuando una chica tiene problemas en una cita no debe culpar a nadie sino a sí misma.
>
> Si no bebes con la pandilla, te hacen sentir como un cobarde.
>
> En ocasiones cuento chistes sucios aunque sería mejor que no lo hiciera.
>
> Papá siempre parece demasiado ocupado para jugar conmigo. (Nettler, 1959, p. 682)

El furor resultante ocasionó que el Consejo Escolar de Houston ordenara la quema de las hojas de respuestas de seis pruebas e inventarios que habían sido aplicadas a 5,000 alumnos de noveno grado.

Es comprensible cómo pudo desarrollarse una situación como ésta cuando nos damos cuenta de que el público general no siempre muestra simpatía por el interés científico en investigar la conducta humana. También se ha alegado que algunos reactivos de las pruebas de personalidad, en particular los que tratan sobre sexo, religión y moral, pueden ser ofensivos en lo personal y potencialmente destructivos del carácter de los niños.

Proyecto Camelot

Otro acontecimiento concerniente a las pruebas psicológicas y a la investigación en la ciencia social en general que produjo fuertes reacciones emocionales a mediados de la década de 1960 fue el Proyecto CAMELOT. Este proyecto, financiado por el gobierno de Estados Unidos, fue diseñado para estudiar las causas de la contrarrevolución y la contrainsurgencia en América Latina. Tanto el público de América Latina como algunos congresistas estadounidenses reaccionaron más bien acaloradamente cuando se enteraron del proyecto, precipitando una investigación del Congreso sobre el uso de pruebas psicológicas en el gobierno, la industria y la educación. Un tema que se ventiló a profundidad durante la indagación fue la aplicación a solicitantes de empleo de reactivos en las pruebas de personalidad concernientes al sexo y la religión, tales como (1) Mi vida sexual es satisfactoria; (2) Creo en Dios, y (3) No me llevo muy bien con mis padres. Las audiencias del Congreso no llevaron a la descontinuación de esas pruebas, pero la preocupación política asociada con las audiencias impulsó a los psicólogos y a otros especialistas en evaluación a prestar más atención a la ética de la evaluación psicológica.

El polígrafo y las pruebas de integridad

El robo es un gran problema en los negocios y la industria estadounidenses, y es posible que cada año sean robados billones de dólares en materiales y productos. En consecuencia, los ejecutivos de las corporaciones están alertas a cualquier medio legal que les permita detectar la deshonestidad entre los empleados o solicitantes de empleo. Durante años, el polígrafo (detector de mentiras), que por lo general mide el ritmo cardiaco, la tasa de respiración, la presión sanguínea y cambios en la resistencia de la piel, fue utilizado por los negocios y las organizaciones industriales para identificar entre sus empleados a los ladrones y mentirosos. Sin embargo, en 1988 el Congreso estadounidense aprobó el Acta de Protección contra el Polígrafo para los Empleados que prohíbe la mayoría de los usos del polígrafo en las entrevistas previas al empleo en el gobierno y el sector privado. Posteriormente se introdujeron varias pruebas de lápiz y papel para tratar de determinar la honestidad o la integridad. Algunos estados también han contemplado la prohibición de esas pruebas, aunque una fuerza de tarea de la Asociación Psicológica Estadounidense concluyó que "las pruebas de honestidad, cuando se utilizan de manera apropiada y en conjunto con otros procedimientos de selección, han demostrado niveles útiles de validez como procedimientos de selección" (APA Task Force, 1991, p. 6).

La práctica de aplicar pruebas de integridad en los negocios y la industria sigue siendo controvertida, y existen muchos problemas sin resolver y preguntas sin responder en relación con el constructo de honestidad y la evaluación de la integridad. El asunto sigue siendo analizado a profundidad en la literatura profesional, la cual es de esperar aclarará los problemas y mejorará las cualidades psicométricas de los instrumentos y la sensibilidad social con la que se emplean (Camara y Schneider, 1994, 1995; Lillienfeld, Alliger y Mitchell, 1995; Ones y Viswesvaran, 1998; Ones, Viswesvaran y Schmidt, 1995; Rieke y Guastello, 1995).

Pruebas de personalidad para la selección de empleados

Con referencia al debate sobre la evaluación de la integridad, se presentó una situación concerniente a la aplicación de una prueba de personalidad de verdadero-falso en el caso de *Soroka* contra la *Corporación Dayton-Hudson* (1991). La disputa estuvo relacionada con PsychScreen, un inventario de personalidad desarrollado a partir del MMPI y el CPI que la gerencia de las Tiendas Target había aplicado a los solicitantes del puesto de guardia de seguridad. Este inven-

tario había sido usado antes para detectar solicitantes para puestos relacionados con el cumplimiento de la ley, el control de tráfico aéreo y las plantas de energía nuclear, en los cuales la seguridad es de suma importancia. La defensa legal del demandante argumentó que los siguientes tipos de reactivos eran discriminatorios con respecto a las preferencias religiosas y sexuales:

> Creo en la segunda llegada de Cristo.
>
> Creo que existen un demonio y un infierno después de la vida.
>
> Me siento muy atraído hacia los miembros de mi propio sexo.
>
> Nunca he tolerado ninguna práctica sexual fuera de lo común. (Hager, 1991, p. A-20)

Los abogados de la Corporación Dayton-Hudson argumentaron que dichas preguntas eran efectivas para identificar a personas emocionalmente inestables, de quienes no podía esperarse un desempeño efectivo en la posición de guardia de seguridad. Sin embargo, la corte de apelación concluyó que las preguntas sobre religión y sexo violan el derecho a la privacidad de la persona que busca empleo y, en consecuencia, falló en favor del demandante.

Al apelar esta resolución ante la Suprema Corte de California, la Asociación Psicológica Estadounidense señaló que reactivos como los del PsychScreen no deberían considerarse por separado sino de manera colectiva al evaluar su efectividad para detectar la inestabilidad emocional. Aún así, puede argumentarse que las preguntas que conciernen a las preferencias sexuales y religiosas, las cuales pueden contribuir ligeramente a la predicción del desempeño en el trabajo pero que casi con certeza no son directamente relevantes para el empleo, no tienen lugar en las pruebas de detección para el empleo.

La evaluación de la personalidad también es citada en la legislación federal que trata con las prácticas justas para el empleo. Por ejemplo, en el Acta de Estadounidenses con Discapacidades (ADA) de 1990 se cuestionó el papel de los inventarios de personalidad y las técnicas proyectivas en el proceso de selección de empleados. De acuerdo con las disposiciones de esta ley, cuando los resultados de las pruebas de empleo se presenten en términos de etiquetas de diagnóstico como "depresión" o "ansiedad", se considera que las pruebas son procedimientos médicos. En consecuencia, de acuerdo con el ADA, las pruebas no pueden ser aplicadas hasta que se haya hecho una oferta condicional de empleo (U. S. Equal Employment Opportunity Commission, 1994). Las pruebas de habilidades cognoscitivas y estados fisiológicos también pueden considerarse parte de un examen médico y, por ende, están sujetas a las mismas restricciones. Sin embargo, en este proceso hay unas cuantas reglas generales y las decisiones se toman caso por caso.

Validez de las pruebas de personalidad

Las preguntas de qué miden las pruebas de personalidad, de si vale la pena medir esas variables, y de cómo interpretar y aplicar mejor los resultados, han recibido un considerable escrutinio en las últimas décadas. Las cualidades psicométricas de los instrumentos de evaluación de la personalidad, y en particular de las técnicas proyectivas, a menudo dejan mucho que desear. Los inventarios con codificación de criterios pueden tener una validez mayor que otros instrumentos o procedimientos, pero sus coeficientes de validez a menudo disminuyen de manera marcada a lo largo del tiempo y con las situaciones.

Se necesita mejorar no sólo la confiabilidad y validez de las pruebas de personalidad, sino también sus bases teóricas y los criterios contra los cuales se validan. El modelo de enfermedad de los trastornos mentales y el sistema asociado de clasificación de diagnóstico (DSM-IV) (American Psychiatric Association, 1994), el cual ha influido en el desarrollo de muchos proce-

dimientos de evaluación de la personalidad, en muchos aspectos son ambiguos y poco confiables. Otro asunto que causa preocupación es la mala interpretación de los resultados de las evaluaciones de la personalidad. Los errores de interpretación pueden ocurrir al no considerar la tasa base, o frecuencia del evento (criterio) a predecir. Las malas interpretaciones también resultan de lo que se conoce como "introspección clínica" o "intuición", que a menudo es sólo una colección de estereotipos superficiales, trivialidades o generalizaciones excesivas.

A pesar de la impresionante variedad de técnicas utilizadas en la evaluación de la personalidad, muchas de ellas representan intentos relativamente imperfectos de medir comportamiento y cognición. Por esta razón, deberían considerarse sobre todo como instrumentos o procedimientos de investigación más que como herramientas psicométricas acabadas. Para ser justos, los inventarios de personalidad y las técnicas proyectivas en ocasiones han contribuido a tomar decisiones acertadas de selección. Un ejemplo de ello es el uso del MMPI en el exitoso programa de selección de los Cuerpos de Paz (Hobbs, 1963; Wiggins, 1973). Combinar medidas de habilidades cognoscitivas con medidas de temperamento y motivación puede también incrementar la capacidad de predicción de los criterios de desempeño en el trabajo. Por ejemplo, Gottfredson (1994) sugirió que la selección para el empleo podía mejorarse identificando los elementos menos cognoscitivos del desempeño en el empleo y aplicando medidas de esos elementos (por ejemplo, ciertos rasgos de personalidad) junto con las pruebas de aptitud. Ella sostuvo que dichos predictores no cognoscitivos pueden reducir el impacto adverso de utilizar predictores cognoscitivos solos y, al mismo tiempo, aumentar la validez de esos predictores. Sin embargo, Gottfredson admitió que la contribución hecha por las variables afectivas, por encima de la aportada mediante una batería de pruebas cognoscitivas, en la predicción del desempeño ocupacional probablemente sea bastante pequeña en la mayoría de los casos.

El problema de la validez de las pruebas de personalidad no puede ser resuelto sin una mejor investigación y desarrollo, pero dichos esfuerzos deberían ser emprendidos con una actitud socialmente responsable y con respeto a los derechos de los individuos (vea Messick, 1995). Los usuarios de las pruebas deben poseer también una sólida comprensión de la estadística y otras cuestiones técnicas concernientes a diseño, confiabilidad, validez y normas de la prueba. Incluso así, es importante que quienes aplican exámenes psicológicos lleven registros de sus aciertos y errores y de otros indicadores de éxito y fracaso que se derivan del uso de los resultados de prueba. A la larga, este tipo de información sirve como una verificación de la validez de las pruebas para lograr sus propósitos declarados.

Sesgo étnico y de género

Un asunto relacionado con los problemas éticos y la cuestión de la validez de la prueba es la pregunta de si las pruebas de personalidad están sesgadas en particular en contra de una raza, un género u otros grupos demográficos. Por ejemplo, Gynther (1981) encontró que los perfiles del MMPI de los negros mostraban mayor grado de psicopatía que los de los blancos. Algunos años más tarde, Dahlstrom y Gynther (1986) concluyeron que esas diferencias eran válidas y no consecuencia de un sesgo en el MMPI contra el grupo étnico. No obstante, en la revisión del MMPI se hizo un esfuerzo por eliminar los sesgos de grupo étnico y de género. Las comparaciones subsecuentes de las calificaciones en el MMPI-2 de hombres afroamericanos y angloamericanos encontraron una serie de diferencias significativas entre los dos grupos (Ben-Porath, Shondrick y Stafford, 1995; Frueh, Smith y Libet, 1996).

Si bien en los años recientes ha sido relativamente poca la investigación sistemática sobre las calificaciones obtenidas en las pruebas de personalidad por diferentes grupos étnicos, de cla-

se social o de nacionalidad, la investigación sobre el sesgo de género ha florecido. Una respuesta tradicional a las diferencias de género en las calificaciones de las pruebas ha sido proporcionar normas separadas para hombres y mujeres, pero también se han hecho esfuerzos por elaborar reactivos de prueba que no estén sesgados hacia ningún sexo. Dichos esfuerzos son rutinarios en la elaboración de pruebas de habilidad que se revisan de manera periódica, como el SAT y el GRE. En lo que respecta a los instrumentos afectivos, el desarrollo del MMPI-2 y de la edición de 1994 del Inventario de Intereses de Strong, en particular, supuso esfuerzos cuidadosos por eliminar el sesgo de género.

Predicción clínica y estadística

El enfoque estadístico (o actuarial) para la obtención de datos y la predicción de la conducta consiste en aplicar una fórmula estadística, un conjunto de reglas o una tabla actuarial a los datos provenientes de la evaluación. Esto puede ser realizado por una persona o, lo que se ha vuelto una práctica común en los años recientes, por una computadora que siga un programa interpretativo. En contraste, el enfoque clínico, o impresionista, supone formular juicios intuitivos o conclusiones basadas en impresiones subjetivas combinadas con una teoría de la personalidad. Las interpretaciones impresionistas no sólo se elaboran sobre la base de entrevistas, datos biográficos y otra información clínica; también pueden usarse calificaciones de personalidad, calificaciones de las pruebas y otros datos basados en la estadística.

Una primera revisión de investigación que compara los enfoques clínico y estadístico hacia la predicción concluyó que en 19 de 20 estudios examinados, el enfoque estadístico era superior o de igual efectividad que el enfoque clínico (Meehl, 1954). Once años más tarde, después de resumir los datos de 50 estudios en los cuales se compararon los dos enfoques, Meehl (1965) concluyó que el enfoque estadístico era más eficiente en dos terceras partes de los estudios e igual de eficiente que el enfoque clínico en la tercera parte restante. Una revisión subsecuente de Sines (1970) coincidió con la conclusión de Meehl: en todos salvo uno de los 50 estudios revisados por Sines se encontró que el enfoque actuarial (estadístico) era superior al clínico en la predicción de varios tipos de conducta.

Aunque los estudios resumidos por Meehl y Sines dieron un apoyo impresionante a la conclusión de que los diagnósticos de personalidad y las predicciones de conducta tienen mayor precisión cuando se emplea un enfoque estadístico que cuando se usa uno clínico, Lindzey (1965) demostró que un clínico experto puede, en ocasiones, formular diagnósticos de gran precisión. Usando sólo la información obtenida de la aplicación del Test de Apercepción Temática, un psicólogo clínico demostró una exactitud de 95% para detectar la homosexualidad. El enfoque estadístico de emplear sólo ciertas calificaciones objetivas obtenidas de los protocolos del TAT fue significativamente menos preciso.

Otros estudios también han encontrado que, en ciertas circunstancias, profesionales entrenados que emplean datos de una variedad de fuentes (historia de caso, entrevista, batería de pruebas, etc.) hacen mejores predicciones que las fórmulas actuariales (por ejemplo, Goldberg, 1970; Holt, 1970; Wiggins y Kohen, 1971). El debate sobre la efectividad relativa de los enfoques clínico y estadístico hacia la evaluación de la personalidad ha disminuido, pero la investigación sobre el tema continúa. Por ejemplo, Gardner, Lidz, Mulvey y Shaw (1996) compararon la exactitud de un procedimiento actuarial con la de uno clínico en la predicción de la conducta violenta de personas con enfermedad mental. Los pacientes fueron seguidos durante seis meses desde su liberación en la comunidad después de haber sido vistos en una sala de emergencias psiquiátrica. Como en la gran mayoría de las comparaciones anteriores de los enfoques clínico

y actuarial (estadístico) hacia la predicción, el enfoque estadístico demostró ser superior al clínico en una serie de criterios. Las tasas de errores por falso positivo y falso negativo fueron menores en el enfoque actuarial. Las predicciones actuariales basadas sólo en la historia de violencia de los pacientes fueron más exactas que las predicciones clínicas, y las predicciones actuariales que no usaron información sobre las historias de los pacientes también fueron más precisas que las clínicas.

Rasgos y situaciones

El énfasis en las situaciones, en oposición a los rasgos, como determinantes de la conducta se remonta a los estudios de Hartshorne y May (1928) sobre el carácter de los niños. Cuatro décadas más tarde, Walter Mischel (1968) elaboró un resumen con evidencia suficiente para apoyar la conclusión de que, aunque los correlatos conductuales de las habilidades cognoscitivas son bastante consistentes entre diferentes situaciones, la conducta personal-social depende en gran medida de la situación específica. Mischel concluyó que las inferencias relativas a la dinámica o los rasgos de personalidad son menos útiles que el conocimiento de la situación en sí misma para predecir la conducta. Argumentó que las evaluaciones de rasgos generalizados de personalidad no son de particular utilidad porque dichos rasgos con frecuencia no se generalizan entre situaciones. En lugar de analizar la personalidad en un complejo de rasgos o factores, Mischel (1986) propuso un enfoque del aprendizaje social. Este enfoque enfatiza que la gente aprende a dar respuestas diferentes en situaciones distintas, y que la precisión con la que puede predecirse la conducta de una persona en un contexto situacional específico debe tener en consideración la historia de aprendizaje de esa persona en situaciones similares.

Es cierto que las normas sociales, los roles y otras condiciones relacionadas con el grupo ejercen poderosos efectos en la gente y pueden anular al temperamento o estilo personal como determinantes de los pensamientos y acciones del individuo. Cuando una situación social permanece totalmente constante, las personas tienden a suprimir sus idiosincracias y a adaptar su conducta y sus pensamientos a las expectativas sociales, recompensas y castigos proporcionados en esta situación. La investigación en psicología social y algunos programas de televisión han demostrado ampliamente que todos los tipos de personas siguen el dicho “a donde fueres haz lo que vieres”. Sin embargo, la aceptación de este lugar común no significa que la personalidad individual no juegue algún papel en la determinación de la conducta.

No toda la evidencia referente a la controversia rasgo-situación favorece al situacionismo. Algunos investigadores (por ejemplo, Bem y Allen, 1974; Block, 1977; Chaplin y Goldberg, 1984; Underwood y Moore, 1981) han encontrado que la consistencia de los rasgos a través de las situaciones es en sí misma una diferencia individual variable. Independientemente de las circunstancias externas, algunas personas son más consistentes que otras en su comportamiento, y por lo general son conscientes de la consistencia de su comportamiento. En una investigación de Bem y Allen (1974), las personas que creían ser muy consistentes en la amistad y en la rectitud tendían a serlo; quienes se identificaban como menos consistentes tendían a ser de esa manera.

La investigación también ha demostrado que algunos comportamientos son más consistentes que otros. Ciertas conductas son estrechamente específicas a la situación, mientras que otras que no requieren de estímulos provocadores específicos ocurren en una amplia gama de situaciones y, por ende, reflejan mejor las variables amplias de personalidad (Funder y Colvin, 1991).

Al revisar la posición de Mischel y la evidencia subsecuente, una conclusión razonable es que existe poco apoyo para un punto de vista situacionista estricto concerniente a la personalidad. Más bien, es mejor enfatizar que la conducta es un producto conjunto de las características

de personalidad y la situación particular en la cual ocurre. En ciertas situaciones (fuertes), los propios rasgos de las situaciones son más importantes en la determinación de cómo se comporta la gente; en otras situaciones (débiles), las características de personalidad tienen más influencia. Las personas con ciertos rasgos de personalidad también tienden a buscar situaciones que tienen ciertas características. La gente no sólo es afectada por las situaciones específicas, sino que en cierta medida elige las situaciones que la afectarán.

Enfoques idiográfico y nomotético

Al igual que los debates sobre la precisión relativa de los enfoques clínico y estadístico a la recopilación y análisis de los datos de evaluación de la personalidad, y sobre la relativa importancia de los rasgos y las situaciones en la determinación de la conducta, la controversia entre las posiciones idiográfica y nomotética se ha acallado con el paso del tiempo. Tal como lo planteó Allport (1937), el *enfoque idiográfico* sostiene que cada persona es un sistema legal e integrado que debe estudiarse como un individuo por derecho propio. En el *enfoque nomotético*, que se basa firmemente en normas de grupo o promedios, se buscan y administran leyes generales de la personalidad y la conducta que puedan aplicarse a todas las personas. En lugar de tratar de interpretar en relación con las normas las calificaciones obtenidas por una persona en un conjunto de pruebas, inventarios, escalas de calificación e instrumentos estandarizados similares, los psicólogos que tienen una convicción idiográfica estudian las consistencias y variaciones en la persona, tal como se advierte en observaciones, entrevistas y registros personales (diarios, biografías y escritos similares).

RESUMEN

Las características temperamentales, emocionales y de estilo referidas como variables de personalidad no son tan estables ni se miden con tanta precisión como las variables cognoscitivas. Los intentos por evaluar esas características se remontan a la antigüedad, pero no fue sino hasta finales del siglo XIX y principios del XX que se puso en marcha una aproximación científica genuina a la evaluación de la personalidad.

Aunque algunos instrumentos de evaluación han sido diseñados sobre una base puramente empírica, muchos han sido elaborados en el contexto de una teoría de la personalidad. A este respecto, las teorías psicoanalítica, de rasgo-factor y fenomenológica han ejercido particular influencia. Más recientemente, la teoría del aprendizaje social también ha estimulado el desarrollo de una serie de instrumentos y procedimientos para evaluar conductas características.

Los psicólogos clínicos aplican pruebas y otros instrumentos psicométricos para detección, psicodiagnóstico, planeación de tratamientos e investigación en las clínicas de salud mental y en otros escenarios. De particular importancia son los exámenes del estado mental, los cuales evalúan el estado intelectual, perceptual-motriz y emocional de los pacientes por medio de entrevistas a profundidad, cuestionarios, escalas de calificación y procedimientos psicométricos relacionados. Después de completar un examen psicodiagnóstico de una persona, se realiza una conferencia de caso clínico para explicar los resultados a los miembros de la familia y a otras personas que tienen el derecho de conocerlos.

Los psicólogos de la salud y los psicólogos legales son entrenados para realizar una variedad de tareas en contextos médicos o de cumplimiento de la ley. Los psicólogos de la salud analizan el papel de los factores psicológicos en la enfermedad física y ayudan a planear y a poner

en práctica tratamientos prescritos para tales condiciones. Entre las muchas actividades de los psicólogos legales o forenses está la evaluación psicológica tanto de los transgresores como de las otras partes involucradas en casos judiciales relacionados con cuestiones de competencia para enfrentar un juicio, responsabilidad por actos delictivos y la custodia de menores.

La validez de los instrumentos de evaluación de la personalidad y la forma en que ésta varía con la cultura, el grupo étnico y el género, son y seguirán siendo un asunto de interés. Se han realizado esfuerzos por eliminar, o al menos controlar, los sesgos de género en los inventarios de intereses y en otras medidas afectivas reescribiendo los reactivos con el propósito de volverlos relevantes y justos para hombres y mujeres, y proporcionando normas separadas y combinadas para ambos sexos. Sin embargo, la elaboración de instrumentos justos para el género y la raza no reduce la necesidad de medir y estudiar las diferencias individuales y de grupo en las características de personalidad.

Las pruebas de personalidad y otras medidas afectivas, en especial cuando se aplican en situaciones educativas y de empleo, han sido criticadas por representar invasiones a la privacidad, por ser irrelevantes o hacer malas predicciones de la conducta, e incluso por sugerir actos inmorales. Los psicólogos reconocen las limitaciones de esos tipos de instrumentos y, por regla general, las críticas han tenido el efecto saludable de incrementar el interés en el diseño de nuevos instrumentos de evaluación afectiva.

En los años recientes ha disminuido también el debate sobre otros temas concernientes a la evaluación de la personalidad, como la efectividad relativa de la predicción clínica y estadística, la importancia relativa de los rasgos físicos y las situaciones como determinantes de la conducta, y el valor relativo de examinar la unicidad del individuo (enfoque idiográfico) en oposición a la búsqueda de leyes generales de la conducta que se apliquen a toda la gente (enfoque nomotético).

PREGUNTAS Y ACTIVIDADES

1. Describa los conceptos principales de las siguientes teorías: de los rasgos, psicoanalítica, fenomenológica (del yo) y del aprendizaje social. ¿Qué teorías han contribuido más significativamente a la evaluación de la personalidad? ¿Cuál teoría resulta más atractiva para usted en términos de su poder explicativo y de congruencia con su propia teoría de la personalidad humana?

2. Defienda la fisionomía y la grafología como áreas legítimas de investigación y aplicación en la evaluación de la personalidad. ¿Por qué son más respetables que la frenología y la astrología?

3. Muestre a varios de sus amigos la descripción de la personalidad que se presenta en la página 323. ¿Cuántos están de acuerdo en que es una descripción bastante precisa de su personalidad? ¿A qué atribuye esos resultados?

4. Consulte en la sección amarilla del directorio telefónico de varias ciudades grandes, que puede encontrar en la mayoría de las bibliotecas públicas, los anuncios acerca de servicios psicológicos. Busque en varios apartados, incluyendo psicólogos, psiquiatras, psicoterapeutas, médicos, consejeros, terapeutas, consejeros matrimoniales, educación, clínicas o cualquier otro concepto que a usted se le ocurra pueda ser relevante. ¿Qué información se da para ayudar a la gente que busca dichos servicios? Además del directorio telefónico, la Sociedad Médica del Condado o región, el Centro de Salud Mental y otras organizaciones locales deben poder brindarle un listado de quienes proporcionan servicios psicológicos.

5. Distinga entre los conceptos legales de competencia y demencia. ¿Qué instrumentos o técnicas psicológicas de evaluación pueden contribuir a tomar decisiones concernientes a la competencia o la demencia?

6. Describa varios papeles que desempeñan los psicólogos en los escenarios de cumplimiento de la ley. ¿Cuáles de esos papeles son más válidos y socialmente útiles?

7. Muchos artículos y programas en los medios de comunicación han tratado con el problema del abuso infantil y los procedimientos para detectar y confirmar si éste ha ocurrido en casos específicos. Realice una búsqueda en la biblioteca y en fuentes de Internet sobre los materiales y procedimientos usados por los psicólogos para determinar si se ha abusado de un niño. ¿Qué tan válidas son esas técnicas y cuáles son sus peligros y otros defectos?

8. ¿Qué es predicción clínica y cómo difiere de la predicción estadística? ¿Cuál es más efectiva y por qué?

9. Escriba reactivos del tipo falso-verdadero para una prueba de personalidad que estén sesgados hacia los hombres, las mujeres, los negros, los blancos, los asiáticos y las personas con mayor educación.

CAPÍTULO QUINCE

OBSERVACIONES Y ENTREVISTAS

Existen muchas formas de obtener información acerca de la personalidad, de las cuales las más populares se consideran en los siguientes cuatro capítulos. Algunos de esos enfoques a la evaluación de la personalidad se basan en alguna teoría; otros son más empíricos u orientados a los hechos. Algunas formas son más directas, otras son indirectas o incluso intrincadas. Algunas son complejas y costosas; otras sencillas y económicas. Este capítulo trata acrca de los procedimientos de evaluación de la personalidad que pueden considerarse los más empíricos y directos, pero no por necesidad los menos costosos: la observación y la entrevista. Estos procedimientos, que pueden emplearse como fuentes primarias o secundarias de información acerca de la gente y para propósitos distintos a la evaluación de la personalidad, implican los actos, al parecer superficiales, pero a menudo complicados, de observar y escuchar. Cuando son empleados por individuos astutos, pero considerados, pueden proporcionar un caudal de información sobre las formas típicas de actuar y pensar de una persona. Por otro lado, cuando el observador o el entrevistador no es experimentado, es poco sensible o es tendencioso, los resultados pueden ser engañosos o incluso inútiles.

OBSERVACIONES

El método de evaluación de la personalidad empleado de manera más amplia y tal vez el que mejor se entiende y acepta es alguna forma de observación. La observación, que es básica para todas las ciencias, consiste en que el observador simplemente tome nota de ciertos acontecimientos, como una conducta particular, y por lo general lleve un registro de lo que observa. El procedimiento más común es la *observación no controlada* de la conducta "sobre la marcha", sin tratar de restringirla a una situación o conjunto de circunstancias particulares. La observación de las actividades de los niños en un patio de juegos y de la conducta de la gente en una fila de espera son ejemplos de esta observación no controlada, o *naturalista*. Un ejemplo de observación no controlada en el mundo laboral es la *técnica de incidentes cruciales* (Flanagan, 1954). Se pide a los supervisores y a otras personas familiarizadas con cierto trabajo que identifiquen conductas específicas que son cruciales para el desempeño o que distingan entre buenos y malos trabajadores en el puesto. Esas conductas, o *incidentes*, son cruciales porque tienen consecuencias muy positivas o muy negativas. Algunos ejemplos son "asegurar la maquinaria y limpiar el lugar de trabajo al terminar" y "dar un pronto seguimiento a las solicitudes de los clientes". La identificación de un gran número de dichos incidentes proporciona información valiosa sobre la naturaleza del empleo y los requisitos para desempeñarlo de manera efectiva.

Las observaciones pueden ser no controladas y aun así ser sistemáticas y objetivas. Por ejemplo, puede entrenarse a los profesores para efectuar observaciones objetivas del comporta-

miento de los escolares y *registros anecdóticos* precisos de cualquier conducta que parezca importante. Un maestro-observador bien entrenado indica con precisión en el registro anecdótico lo que se observó y lo distingue de la forma en que se interpretó. El observador se da cuenta de que cuando Juanito pellizca a María no siempre es un acto de agresión.

Mejoramiento de la precisión de las observaciones

Una de las directrices recomendadas para mejorar la validez de los datos observacionales es entrenar a los observadores para que sean tan objetivos como sea posible, sin permitir que sus sesgos y necesidades personales afecten lo que observan y puedan separar la observación de la interpretación. Otra directriz es observar un número limitado de conductas específicas, las cuales se definen de antemano. El empleo de varios observadores y la obtención de una muestra grande y representativa de observaciones también puede mejorar la precisión de éstas. Sin embargo, la obtención de una muestra representativa de la conducta consume tiempo y es costosa. Para reducir el volumen de datos obtenidos en la observación continua resulta apropiada la técnica de *muestreo de incidentes*, la cual consiste en advertir y registrar sólo eventos o incidentes específicos, por ejemplo, de conducta agresiva. Es posible obtener otras mejoras en la eficiencia de la observación mediante el *muestreo de tiempo* —efectuar una serie de observaciones que duran sólo unos cuantos minutos a lo largo de un día o algo por el estilo (Wright, 1960).

Observación participante

La *observación participante* es también relativamente no controlada, en ésta el observador forma parte de la situación que se observa. La observación participante ha sido usada de manera general por los antropólogos culturales, tanto que en una época se decía que una familia aborigen típica constaba de una madre, un padre, dos hijos y ¡un antropólogo cultural! Al señalarse el deber de tener en cuenta la probabilidad de que la propia conducta del observador afectara las reacciones de las otras personas involucradas en la situación, los defensores de este método argumentaron que la participación activa en una situación puede proporcionar introspecciones que no pueden obtenerse por otros medios.

Pruebas de situación

Además de las observaciones relativamente no controladas, se realizan observaciones convenidas con anterioridad, artificiales o controladas, con el propósito de determinar cómo se comportan las personas (y los animales) en varias situaciones. Por ejemplo, un psicólogo del desarrollo puede establecer de antemano una situación de observación para determinar si los niños harán trampa o se comportarán con honestidad en un conjunto de circunstancias arregladas previamente. O al observar la conducta de los niños en una situación similar al juego que incluye muñecos u otros juguetes, un observador puede obtener evidencias para confirmar o descartar que son víctimas de abuso.[1]

Una serie clásica de estudios que utilizaron procedimientos controlados de observación conocidos como *prueba de situación* fue la Encuesta de la Educación del Carácter (Hartshorne

[1]Ejemplo de una prueba de situación para evaluar los programas sobre abuso sexual infantil es la Prueba de Situaciones "Qué pasaría si" (WIST), la cual fue diseñada con el propósito de evaluar las habilidades de los preescolares para reconocer, resistir e informar de contactos inadecuados.

y May, 1928). En estas investigaciones se brindó subrepticiamente a los niños la oportunidad de demostrar su honestidad, altruismo y otros rasgos de carácter. Por ejemplo, para probar la honestidad los investigadores colocaron a los niños en una situación donde podían robar algunas monedas o en otra donde podían copiar las respuestas de un examen, supuestamente sin ser detectados. Entre otras cosas, los estudios encontraron que los niños mayores, los menos inteligentes, los de menor posición socioeconómica y los de menor estabilidad emocional tendían a ser menos honestos en todas las situaciones. Quizá el resultado más importante de los estudios de Hartshorne y May fue que la honestidad y otros rasgos de carácter variaban tanto con la situación específica como con el individuo. En otras palabras, el grado de honestidad, altruismo u otras conductas éticas manifestadas por los niños dependía en gran medida de las situaciones en que fueran observados.

Las pruebas de situación para el personal militar fueron introducidas por los alemanes y luego adaptadas por las fuerzas armadas británicas y estadounidenses durante la Segunda Guerra Mundial. La Oficina de Servicios Estratégicos (OSS) de Estados Unidos, precursora de la CIA, diseñó una serie de pruebas de situación simuladas para seleccionar agentes de espionaje. Como en los estudios de Hartshorne y May (1928), se implicó el engaño de los candidatos. Por ejemplo, en el "problema de la pared" se asignó a un grupo de hombres la tarea de cruzar un cañón. Los candidatos reales no sabían que los hombres designados para ayudarles no eran verdaderos candidatos sino que habían sido (insertados) *plantados*. Uno de los plantados actuaba como obstructor haciendo sugerencias poco realistas y comentarios insultantes o fastidiosos; otro plantado simulaba no entender la tarea y resistía pasivamente las instrucciones del candidato. Sin darse cuenta de que los otros candidatos estaban cooperando con los examinadores, el verdadero candidato era observado durante sus esfuerzos por completar la tarea mientras enfrentaba esas circunstancias frustrantes. Sin embargo, fue difícil determinar la efectividad de esos procedimientos como métodos de selección, y nunca se validaron de manera adecuada.

Las pruebas de situación se han usado en otros programas de evaluación, por ejemplo, en la selección de psicólogos clínicos (Kelly y Fiske, 1951). Una variación interesante es la Prueba de Discusión en Grupo sin Líder (LGD), en la cual varios candidatos a un puesto ejecutivo discuten un tema asignado durante 30 a 50 minutos mientras se observa y califica su desempeño individual. Las calificaciones dadas por los observadores, así como por los otros candidatos, pueden ser en términos del grado de dominio, facilitación de la tarea y sociabilidad mostrados por cada uno de los candidatos. A pesar de la calidad realista de las pruebas de situación, nunca es posible duplicar las situaciones reales que los examinados pueden enfrentar. Además, con frecuencia los candidatos se dan cuenta del engaño. Incluso en el programa de evaluación del OSS, algunos candidatos se percataron de que las pruebas estaban arregladas.

Debido en gran medida a lo engañoso de las pruebas de situación y a los problemas para arreglar las situaciones y evaluar los resultados de manera objetiva y consistente, la confiabilidad y la validez predictiva de esas pruebas con frecuencia son demasiado bajas como para justificar el costo.

La conducta de un examinado durante las pruebas de situación puede observarse a través de un monitor de televisión. Al permanecer sin ser visto, el observador no importuna ni afecta la conducta de la gente a la que se observa. Cuando la gente se percata de que está siendo observada puede comportarse de manera no natural o actuar como si estuviera en un escenario (representando un papel). Por esta razón, las observaciones con el propósito de evaluación de la personalidad se efectúan, por lo general, de la manera más discreta posible. En la *observación discreta* el sujeto no está al tanto de la presencia del observador y, por ende, su conducta no es influida por el hecho de saber que está siendo observado. La observación controlada o no con-

trolada puede ser discreta, e incluso la observación participante puede ser relativamente discreta cuando el observador toma medidas para ser aceptado por quienes estén siendo observados.

Observaciones clínicas

Un psicólogo clínico o escolar que examina a un niño interactúa con éste como una clase especial de observador participante. En consecuencia, los examinadores psicológicos deben tener cuidado de no permitir que su presencia y sus acciones provoquen conductas atípicas en el niño. Las observaciones del examinador, las cuales deben ser tan discretas como sea posible, son una parte importante del informe psicológico. Las observaciones deben comunicarse como conductas objetivas y que se puedan verificar de manera que, en lugar de ser expresadas solamente en terminología psicológica, no puedan significar cosas distintas para lectores diferentes.

Mucho de lo que se sabe acerca de la dinámica de la personalidad y los trastornos mentales se ha obtenido de observaciones de la gente en escenarios clínicos. Es obvio que las observaciones clínicas no son del todo objetivas: en una situación clínica cada parte afecta la conducta de la otra. En consecuencia, la precisión de las observaciones clínicas y las interpretaciones que se hacen de ellas deben ser verificadas por otras personas y con procedimientos distintos.

Un observador clínico alerta advierte una variedad de detalles: cómo viste el examinado y si está bien arreglado; si el examinado estrecha la mano del examinador, si lo mira y cómo lo hace; cómo se sienta, se para y camina el examinado; qué expresiones faciales, movimientos corporales y tonos de voz son característicos. Éstas son conductas no verbales y cuando se interpretan de manera apropiada pueden proporcionar una mejor información sobre la personalidad que un registro circunscrito a lo que el examinado dice en realidad.

Entrenamiento de los observadores

Entrenar a los observadores para que sean tan astutos y objetivos como sea posible es más importante que los procedimientos e instrumentos especiales para asegurar la precisión de las observaciones. Debido a que filtran sus observaciones a través de sus tendencias y deseos personales, los observadores que no son sensibles a este hecho a menudo tienen mucha dificultad para efectuar observaciones precisas y separar la observación de la interpretación o el hecho de la opinión.

El entrenamiento de los observadores empieza por describir la forma o el programa base para efectuar las observaciones y revisar la definición objetiva de cada conducta meta y cómo se van a registrar su ocurrencia y duración. Debe decirse a los observadores qué buscar y cómo registrar sus observaciones de manera clara, objetiva y discreta; cómo distinguir entre lo que se observa y la manera en que se interpreta, y cómo estar más al tanto de los efectos de sus tendencias personales y otros factores en lo que observan e informan. El entrenador señala los errores comunes cometidos al registrar las conductas y la importancia de no permitir que los sesgos, expectativas, personalidad, actitudes o deseos personales interfieran con lo que está siendo observado.

Dado que el conocimiento previo acerca de cierta gente puede dar lugar a suposiciones o expectativas de comportamientos típicos, a los observadores sólo se les debe proporcionar la información absolutamente esencial sobre las personas a las que van a observar. Para minimizar el sesgo en las observaciones creado por el deseo de proporcionar al investigador o supervisor datos que lo apoyen, los observadores deben recibir información mínima concerniente a los propó-

sitos del proyecto de investigación y no se les debe dar detalles acerca de las hipótesis específicas o los resultados esperados. Siempre que sean visibles para las personas observadas, debe advertirse a los observadores que se vuelvan lo menos notorios y lo más discretos posible, que permanezcan en el fondo y registren lo que ven y escuchan sin mostrar emoción, aprobación o desaprobación. A las personas que se entrena para ser observadores también se les debe dar la oportunidad de practicar o representar sus actividades de observación y recibir la evaluación de su desempeño antes de hacer observaciones genuinas. Para asegurar la confiabilidad elevada de las observaciones es preferible contar con dos o más observadores a tener uno. También es preferible definir las conductas a observar de manera tan específica como sea posible en lugar de designarlas en categorías descriptivas muy generales.

Conducta no verbal

La mayoría de las personas se da cuenta de que la comunicación interpersonal no es del todo verbal, pero por lo regular no está al tanto de la medida en que los movimientos de sus manos, ojos y boca, así como su postura corporal y tono de voz se interpretan como mensajes. Como se sugiere en la siguiente cita, Sigmund Freud estaba bien consciente de esas señales no verbales: "El que tenga ojos para ver y oídos para escuchar puede convencerse de que ningún mortal puede guardar un secreto. Si sus labios guardan silencio, conversa con las yemas de los dedos; la revelación transpira en cada poro" (Freud, 1905, p. 94).

Se ha realizado una gran cantidad de investigación sobre la conducta no verbal, incluyendo la *cinestesia* (movimiento de las partes del cuerpo), la *proxémica* (distancia entre los comunicantes) y la *paralingüística* (tono de voz, ritmo del habla y otros aspectos no verbales del lenguaje). De acuerdo con los hallazgos de una investigación, de 65 a 90% del significado en las comunicaciones interpersonales proviene de las señales no verbales (Mehrabian y Weiner, 1967).

Se ha encontrado que ciertos tipos de señales no verbales son más importantes que otros en la transmisión de mensajes. Los cinestésicos son particularmente importantes, seguidos de los proxémicos, los paralingüísticos e incluso los *culturales* (estilo de vestir, hábitos o costumbres basados en la cultura, etc.). Es probable que la mayoría de las personas logre más aciertos que errores al interpretar mensajes no verbales, pero los errores ocurren. El rostro del apostador de póker y el estrechamiento efusivo de manos del vendedor o el político son famosos por su habilidad para engañar y manipular a otra gente por medio de la conducta no verbal. Las conductas y las características no verbales se interpretan con mayor precisión cuando el observador tiene algún conocimiento de la situación específica o contexto en el que ocurren. Además, algunas personas son mejores que otras para interpretar la conducta no verbal, una habilidad que parece estar relacionada con la personalidad, pero no con la inteligencia.

El PONS. Rosenthal, Hall, DiMatteo, Rogers y Archer (1979) elaboraron el Perfil de Sensibilidad No Verbal (PONS) para evaluar diferencias individuales en la habilidad para interpretar comunicaciones no verbales. El PONS consiste en una película de 45 minutos en la cual se presenta a los espectadores una serie de estímulos como expresiones faciales o frases habladas escuchadas como tonos o sonidos, pero no como palabras. Después de que se presenta cada estímulo, el espectador selecciona la más apropiada de dos etiquetas descriptivas. Los autores del PONS informan que los hombres y las mujeres que obtienen altas puntuaciones en la prueba tienden a tener menos amigos, pero relaciones sexuales más cálidas, honestas y satisfactorias que quienes obtienen bajas puntuaciones.

Con el razonamiento de que la sensibilidad a los mensajes no verbales es una habilidad importante para los diplomáticos, David McClelland utilizó el PONS en el programa de detección de solicitantes de empleo para la Agencia de Información de Estados Unidos. A estos solicitantes se les presentaron segmentos cortos grabados de la prueba y se les pidió que indicaran qué emoción estaba siendo expresada. Se encontró que quienes calificaron alto en el PONS eran considerados por sus colegas como significativamente más competentes que los que obtuvieron bajas puntuaciones (Rosenthal *et al.,* 1979, pp. 304-306).

Desenmascarar el rostro. Otra contribución a la evaluación de la conducta no verbal es el Sistema de Codificación de la Acción Facial (FACS). Diseñado por Paul Ekman y Wallace Friesen (1978, 1984), el material del FACS consta de 135 fotografías de varias expresiones faciales para entrenar a los observadores en la calificación de docenas de unidades de acción facial. Kaiser y Wehrle (1992) desarrollaron luego un método, basado en el FACS, para la codificación automatizada de la conducta facial en una prueba asistida por computadora o en situaciones de juego. El instrumento Fotografías del Afecto Facial, de Friesen y Ekman, también es útil para enseñar a los observadores a juzgar la emoción a partir de las expresiones faciales. Se trata de 110 fotografías en blanco y negro que expresan temor, enojo, felicidad, tristeza, sorpresa o peligro (además de una expresión neutral).

Autoobservación y análisis de contenido

Mucha gente suele pasar gran cantidad de tiempo observándose, y éste es un método útil para obtener datos observacionales con propósitos clínicos y de investigación. La autoobservación no sólo es un procedimiento económico de investigación, sino una de las pocas formas de tener acceso a los pensamientos y sentimientos privados.[2] Un problema de las autoobservaciones es que probablemente sean aún más tendenciosas que las observaciones realizadas por otros. Es raro que la gente sea del todo objetiva al describir sus pensamientos y su conducta (Wolff y Merrens, 1974). Sin embargo, como con las observaciones realizadas por otros, es posible entrenar a las personas para que efectúen observaciones de sí mismas más objetivas y sistemáticas (Thoreson y Mahoney, 1974). Pueden aprender a distinguir entre lo que realmente sienten, piensan o hacen de lo que deberían o les gustaría sentir, pensar o hacer. Por ejemplo, los hallazgos de la investigación muestran que tener la oportunidad de "vernos como nos ven los otros" puede lograr que nuestras autopercepciones y autoevaluaciones sean más parecidas a las de la demás gente. Por ejemplo, Albright y Malloy (1999) confirmaron la hipótesis de que el observarse en una cinta de vídeo en interacción social incrementa la precisión con que se pronostican los juicios que otros emiten sobre uno mismo.

Al llevar un registro continuo por escrito de pensamientos, sentimientos y acciones puede acumularse un caudal de datos de autoobservación. Por desgracia, no siempre está claro qué hacer con esa abundancia de datos, es decir, cómo analizarlos o interpretarlos. Como se ve en el *análisis del contenido* de diarios, autobiografías, cartas, dibujos y otros documentos personales, es posible obtener información importante sobre la personalidad y la conducta al interpretar los datos de la autoobservación (Allport, 1965). Pero la complejidad y laboriosidad del análisis del contenido ha impedido que este enfoque interpretativo se aplique de manera rutinaria en la clínica y otros contextos aplicados. Wrightsman (1994) presenta una breve revisión de los usos psicológicos y defectos de varios tipos de documentos personales y la autobiografía en particular.

[2]Otros son la hipnosis, el narcoanálisis y la asociación libre.

DATOS BIOGRÁFICOS

Psicobiografía

Además de sus usos potenciales en el diagnóstico clínico, los documentos personales como diarios, cartas y autobiografías proporcionan una fuente rica de información para los psicobiógrafos. La *psicobiografía* es una subcategoría de la psicohistoria: ambas emplean conceptos y teorías psicológicas para reconstruir e interpretar lo ocurrido en el pasado. De manera más específica, la *psicohistoria* se interesa en el análisis, por medio de la historia y la psicología, de sucesos como los juicios por brujería en Salem o el ascenso de la Alemania nazi. Por otro lado, el término *psicobiografía* se refiere a la exploración psicológica de la vida de una persona (Wrightsman, 1994).

Los estudios psicobiográficos de muchas personas famosas, incluyendo a líderes políticos como Adolfo Hitler (Binion, 1976; Langer, 1972), Mohandas Gandhi (Erikson, 1969) y varios presidentes estadounidenses (Brodie, 1983; Freud y Bullitt, 1967; Glad, 1980; Kearns, 1976; Mazlish, 1973), se han conducido con propósitos teóricos y prácticos. Entre las razones prácticas están proporcionar a los líderes de oposición o a otros que deben tratar con ciertas figuras políticas, información sobre la personalidad y la conducta de esos líderes y predicciones de lo que harían en ciertas circunstancias. Esos fueron los motivos detrás de la psicobiografía que Freud y Bullitt (1967) efectuaron de Woodrow Wilson y la psicobiografía que Langer (1972) realizó de Adolfo Hitler.

La psicobiografía ha sido criticada por varios errores factuales, teóricos, culturales y lógicos. La revisión de Wrightsman es muy crítica de este enfoque, pero algunos críticos como Elms (1976) y Cocks y Crosby (1987) han emitido sugerencias para mejorar los procedimientos psicobiográficos. Argumentan que no debe efectuarse un análisis psicobiográfico a menos que se disponga de suficiente información sobre la vida de la persona o sobre las áreas o periodos que se estén analizando. Además, deben aplicarse al análisis otras teorías psicológicas aparte del psicoanálisis clásico. Por último, las ideas preconcebidas y los sesgos de los psicobiógrafos deben reconocerse y controlarse.

Además de las autobiografías y otros documentos personales, la información biográfica registrada en formularios de solicitud, cartas de recomendación y respuestas dadas en inventarios biográficos (datos biográficos) puede contribuir a profundizar en el conocimiento de características de personalidad. Esas fuentes se usan de manera extensiva en las decisiones de empleo y admisión, pero también pueden demostrar ser valiosas en la evaluación de la personalidad y en el diagnóstico de trastornos conductuales y sus causas (vea Stokes, Mumford y Owens, 1994).

Datos biográficos en los contextos de empleo

Los datos biográficos que conciernen a las características, experiencias y logros de una persona se basan en las observaciones de la propia persona, así como en los de otra gente. Obtenida por lo regular a partir de formularios de solicitudes y de otros formularios de autorreporte, la información autobiográfica es útil a los propósitos de toma de decisiones en los contextos educativo, médico, recreativo, de empleo, etc. Sin embargo, la investigación y las aplicaciones más sistemáticas con los datos autobiográficos han ocurrido en situaciones de empleo. Aunque gran parte de esta información se basa en hechos y es objetiva (nombre del solicitante, fecha de nacimiento, estado civil, etc.), una cantidad importante se obtiene de autoobservaciones y de las impresiones del sujeto sobre el ambiente interpersonal.

Solicitudes y recomendaciones. Entre las primeras cosas que se requieren de un solicitante de empleo está el hacer una carta de solicitud y/o llenar un formulario de solicitud. Un formulario de solicitud lleno es un requisito formal para el empleo y una breve descripción de la aptitud del solicitante para el puesto. Luego de una serie de preguntas de identificación (nombre, dirección, empleo deseado, etc.), se solicita información antecedente detallada (educación, impedimentos físicos, registro militar, empleos y experiencia previos). En la mayoría de los casos, se proporciona una sección del formulario para referencias.

Sea obtenida por carta, teléfono, entrevista o cuestionario, la información de las referencias mencionadas por un solicitante puede ser útil a pesar de ciertas limitaciones obvias. Es probable que la limitación más seria de las cartas de recomendación sea que a menudo proporcionan una descripción sesgada o tendenciosa del solicitante. En efecto, el elogio es tan común en las cartas de recomendación que los administradores de personal y otros encargados de la selección con frecuencia se sensibilizan mucho a cualquier cosa que no sea una afirmación muy positiva acerca del solicitante. También existe una tendencia a interpretar las cartas breves como indicativas de desaprobación y las cartas largas como más elogiosas. Debido a que los antiguos empleadores y otras fuentes de referencia se muestran renuentes a revelar por escrito información negativa acerca de una persona, en ocasiones una llamada telefónica vale por una docena de cartas de recomendación.

Inventarios biográficos. Los inventarios biográficos formales, o formas de datos biográficos, constan de una variedad de reactivos que atañen a la historia de vida de un solicitante (relaciones familiares, amistades, actividades extracurriculares, intereses, etc.). Se ha realizado gran cantidad de investigación sobre formas más extensas de los formularios de solicitud ponderados con empleados en todos los niveles de una organización (Schoenfeldt y Mendoza, 1994; Stokes, Mumford y Owens, 1994).

Los inventarios biográficos no sólo tienen una gran validez de contenido, también pronostican muy bien el desempeño en una variedad de contextos de trabajo que van desde el trabajo que no requiere de muchas habilidades hasta las responsabilidades de alto nivel ejecutivo (Childs y Klimoski, 1986; Drakeley, Herriot y Jones, 1988). En muchos casos la validez de esos inventarios también puede generalizarse de un contexto a otro (Rothstein, Schmidt, Erwin, Owens y Sparks, 1990). A pesar de esas ventajas, los inventarios biográficos no se usan mucho con propósitos de selección de personal (Hammer y Kleiman, 1988). Una explicación es que existen problemas legales asociados con la solicitud de ciertos tipos de información (por ejemplo, edad, sexo, grupo étnico, religión, estado civil, número de hijos) en los formularios de solicitud e inventarios biográficos. Además, los solicitantes pueden objetar ciertos reactivos (finanzas personales, antecedentes familiares y otros detalles íntimos) por ser demasiado personales u ofensivos (Rosenbaum, 1973). Esto es desafortunado porque las respuestas a esos reactivos con frecuencia permiten una buena predicción del desempeño laboral.

ENTREVISTAS

La entrevista es uno de los métodos más antiguos y de uso más frecuente para la evaluación de la personalidad. Una entrevista no sólo arroja el mismo tipo de datos que las observaciones, también proporciona información sobre lo que la persona dice y hace. La conducta no verbal del entrevistado, incluyendo sus posturas y desenvoltura, gestos, movimientos oculares y calidad y

patrón del habla, es importante y debe observarse. Sin embargo, el énfasis principal de la entrevista está en el contenido de las afirmaciones verbales del entrevistado. Por esta razón, una entrevista puede definirse como un "intercambio verbal cara a cara en el cual una persona, el entrevistador, intenta obtener información o expresiones de opinión o creencia de otra persona o personas" (Maccoby y Maccoby, 1954, p. 449). La información obtenida en una entrevista consiste en detalles de los antecedentes o la historia de vida del entrevistado, además de datos concernientes a sus sentimientos, actitudes, percepciones y expectativas.

Las entrevistas se emplean en muchos contextos diferentes y con diversos propósitos. En los contextos de investigación se utilizan para encuestas, estudios y para obtener información a profundidad sobre la personalidad y la conducta con propósitos de probar alguna hipótesis o propuesta teórica. En las situaciones de empleo, las entrevistas se utilizan para la selección y detección de empleados, la evaluación o valoración, resolución de problemas y liquidación. En los contextos clínicos, las *entrevistas de ingreso* de los pacientes y sus familiares son esenciales en la obtención de información de historia de caso para formular diagnósticos médicos y/o psicológicos (*entrevistas de diagnóstico*). Además, las *entrevistas terapéuticas* forman parte del proceso del tratamiento psicológico, y las *entrevistas de salida* están diseñadas para determinar si un individuo institucionalizado está listo para salir.

Cualesquiera que puedan ser el contexto y los propósitos de la entrevista, ésta requiere habilidad y sensibilidad y puede llevarse mucho tiempo y ser muy laboriosa. La entrevista es tanto un arte como una ciencia, y algunos entrevistadores son más efectivos que otros para establecer *rapport* y lograr que los entrevistados se abran. El procedimiento varía según el propósito de la entrevista, pero, como en cualquier situación interpersonal, los resultados dependen de la personalidad y las acciones del entrevistador y el entrevistado. De este modo, la entrevista no es una situación unidireccional de pregunta y respuesta en la cual el entrevistador no es afectado. Casi en todos los casos es una dinámica, un intercambio en dos direcciones en el cual los participantes se influyen mutuamente.

La entrevista puede constituir un fin en sí misma, pero también puede funcionar como un proceso para familiarizarse o conocerse diseñado como introducción a otros procedimientos de evaluación. A la mayoría de los psicólogos clínicos y asesores les agrada la cercanía cara a cara de una entrevista porque les permite sentir los problemas y las características del paciente (cliente). Los psicólogos clínicos, los psicólogos de personal, los asesores laborales y otros profesionales en servicios humanos creen, por lo general, que el tiempo y el gasto de una entrevista están justificados, porque la información personal que se obtiene de esta manera no puede obtenerse por otros medios. Los solicitantes, los asesorados y los pacientes suelen expresar que se sienten más involucrados cuando son entrevistados que cuando simplemente se les pide responder a cuestionarios de lápiz y papel o formularios de solicitud y no se les da oportunidad de comunicar sus problemas, necesidades, opiniones y circunstancias de una manera personal.

Técnicas de entrevista

Una entrevista personal puede tener lugar en cualquier parte, pero es mejor conducirla en un lugar tranquilo, libre de distracciones. Tanto el entrevistador como el entrevistado deben estar cómodamente sentados y uno frente al otro. Como la entrevista es una habilidad interpersonal compleja y hasta cierto grado una función del estilo interpersonal del entrevistador, no resulta fácil enseñar a conducir una entrevista efectiva. Sin embargo, el prestar atención a las siguientes recomendaciones puede mejorar las habilidades para conducir entrevistas.

Los entrevistadores profesionales son generalmente amistosos pero neutrales, demuestran interés pero no se entrometen ni se manifiestan excesivamente al reaccionar ante los entrevistados. Son cálidos y abiertos, aceptan a sus interlocutores por lo que son sin mostrar aprobación o desaprobación. No empiezan dándole primacía a preguntas del tipo "¿Con cuánta frecuencia golpea a su esposa?", y no formulan preguntas que implican cierta respuesta (por ejemplo, "¿Aún lo hace, no es cierto?"). Al dedicar a las preguntas el tiempo adecuado y al variar el contenido de acuerdo con la situación, los buenos entrevistadores son capaces de desarrollar una conversación que fluye de un tema a otro. Las pausas o silencios no les causan incomodidad: le dan al entrevistado tiempo suficiente para responder una pregunta por completo y escuchan la respuesta sin interrumpirlo. Además, prestan atención no sólo a lo que dice el entrevistado, sino también a cómo lo dice. Al darse cuenta de que la conducta del entrevistador (nivel de actividad, cantidad y velocidad del habla, etc.) tiende a ser imitada por el entrevistado, los entrevistadores son pacientes, se sienten cómodos y no lo apresuran. Los entrevistadores experimentados también verifican sus comprensiones, impresiones y percepciones de las respuestas del entrevistado para aclararlas y asegurarse de que no las entendieron mal. Asimismo, pueden hacer preguntas directas para llenar algunos huecos en su comprensión de los entrevistados, pero no son mirones que de manera despiadada indagan o disfrutan la discusión sobre materiales obscenos o altamente emocionales, ni consciente o inconscientemente refuerzan la atención del entrevistado hacia dichos temas.

Aunque las características de los buenos entrevistadores descritas líneas arriba son de aplicación general, las técnicas específicas varían de acuerdo con la orientación teórica del entrevistador (conductual, centrada en el cliente, psicoanalítica, etc.), así como con las metas y el escenario de la entrevista. Edad, sexo, grupo étnico, atractivo, salud, inteligencia, personalidad y otras características del entrevistado y el entrevistador también pueden afectar el proceso y progreso de la entrevista. La mayoría de los entrevistadores fuera de los contextos clínicos, así como muchos clínicos, son bastante eclécticos en su orientación, no siguen una teoría particular de la personalidad sino que aplican conceptos relevantes de una variedad de teorías.

Entrevistas estructuradas contra entrevistas no estructuradas. El grado en que una entrevista es *estructurada* depende sobre todo de sus metas, pero también es importante considerar las características de los participantes. Algunos entrevistados responden mejor a un enfoque relativamente no estructurado y flexible; otros comunican información más relevante cuando el entrevistador sigue una guía de la entrevista y plantea preguntas muy estructuradas. Los entrevistadores también pueden sentirse más cómodos y obtener mayor cantidad de información personal haciendo una serie de preguntas similares a las que se encuentran en los formularios de solicitud o en una forma de historia personal. Los entrevistadores con menor experiencia generalmente encuentran más sencillo manejar una entrevista estructurada, cuyos resultados pueden cuantificarse con facilidad para su análisis. Los entrevistadores experimentados pueden preferir mayor flexibilidad en el contenido y el tiempo de las preguntas de la entrevista, en otras palabras, menos estructura.

Se requieren más habilidad y tiempo para conducir una entrevista no estructurada o flexible en la cual el entrevistador pueda seguir guías de interés o concentrarse en los detalles de mayor importancia. Para lograrlo, el entrevistador anima al entrevistado a que se sienta en libertad de hablar de sus problemas, intereses, conductas o cualquier otra cosa que parezca relevante para las metas de la entrevista. Esas metas afectan también la cantidad de estructura de una entrevista. Cuando se requieren respuestas a un gran número de preguntas específicas, como en una situación de selección para el empleo, resulta apropiado utilizar un enfoque muy estructurado.

Pero si la meta es obtener una imagen profunda de la personalidad o definir la naturaleza de ciertos problemas y sus causas, se requiere menos estructura. Sea muy estructurada o relativamente flexible, la secuencia de las preguntas va, por lo regular, de lo general a lo específico y de los temas menos personales a los más personales. La mayoría de los entrevistadores profesionales son capaces de variar su enfoque de acuerdo con la personalidad del individuo entrevistado y de los objetivos de la entrevista. Empiezan planteando una serie de preguntas flexibles que no representan una amenaza para establecer *rapport* e iniciar la conversación, y luego las preguntas se vuelven más específicas conforme avanza la entrevista.

Temas y preguntas de la entrevista. Las preguntas específicas que se formulan dependen de los propósitos de la entrevista, pero resulta útil planear ésta señalando los temas que van a cubrirse, si no es que las preguntas específicas que van a plantearse. En la tabla 15.1 se presenta un compendio de la información necesaria a registrar en una entrevista sobre la historia personal.

TABLA 15.1 Información a registrar en una entrevista

Datos de identificación: Nombre, edad, sexo, grado escolar, grupo étnico, nacionalidad, domicilio, fecha de nacimiento, estado civil, fecha de la entrevista y datos similares.

Propósitos de la entrevista: Empleo, ingreso psiquiátrico, psicodiagnóstico, resolución de problemas o de crisis, evaluación del desempeño, terminación o salida.

Apariencia física: Vestimenta, arreglo, descripción física (atractivo, rasgos inusuales, etc.), trastornos o discapacidades físicas evidentes o aparentes.

Conducta: Actitudes y emociones (cooperativo, comunicativo o reservado, amistoso u hostil, defensivo, etc.); conducta motriz (activa contra pasiva, postura, modo de andar, porte); nivel de funcionamiento intelectual (brillante, promedio, retrasado según se estima a partir del vocabulario, memoria inmediata y a largo plazo, juicio, pensamiento abstracto); signos de trastorno mental (procesos distorsionados de pensamiento, construcciones extravagantes, bloqueo del pensamiento, etc.; percepciones distorsionadas: delirios, alucinaciones, desorientación en el tiempo o el espacio, etc.; reacciones emocionales inapropiadas o extremas: depresión, manía; manerismos, posturas o expresiones faciales inusuales).

Familia: Padres, hermanos, otros miembros de la familia; grupo sociocultural; actitud(es) hacia los miembros de la familia.

Historia médica: Salud actual, historia de salud, problemas físicos.

Historia de desarrollo: Desarrollo físico, intelectual, de lenguaje, emocional y social; irregularidades o problemas del desarrollo.

Educación y formación: Escuelas a las que se asistió, nivel de desempeño, ajuste a la escuela, planes para continuar la educación y la formación.

Empleo: Naturaleza y número de empleos o posiciones sostenidas, servicio militar (rango y deberes), nivel(es) de desempeño en el trabajo, problemas en el trabajo.

Problemas legales: Arrestos y condenas, naturaleza de las fechorías o delitos.

Historia sexual y matrimonial: Actividades y problemas sexuales, matrimonios, problemas matrimoniales, separaciones y divorcio(s), hijos.

Intereses y actitudes: Pasatiempos, actividades recreativas, actividades sociales y actitud(es) hacia los demás, nivel de autoaceptación y satisfacción, aspiraciones o metas.

Problemas actuales: Detalles de los problemas presentes y planes para resolverlos.

Una entrevista completa de la historia personal, sea que se conduzca en un contexto clínico, de servicio social, de empleo o de investigación, requiere obtener los tipos de información mencionados en esta tabla. No es necesario cubrir todos esos temas en una situación específica: el entrevistador puede concentrarse en las áreas que considere más importantes. En cualquier caso, las preguntas específicas de la entrevista, redactadas en un lenguaje con el que el entrevistado esté familiarizado y se sienta cómodo, pueden desarrollarse a partir de los lineamientos de la tabla 15.1.

Entrevistas clínicas

Las entrevistas clínicas son conducidas con propósitos de ingreso en una dependencia social u hospital mental, las entrevistas de diagnóstico sirven para determinar las causas y correlatos de los problemas de un individuo, y las entrevistas terapéuticas (consejo, psicoterapia) se dirigen a brindar ayuda. La tabla 15.2 es una lista de recomendaciones a seguir cuando se realiza una entrevista clínica. Muchas de esas recomendaciones no se restringen a las entrevistas clínicas, sino que se aplican también a otros tipos de intercambios verbales.

Cuando se conduce de manera apropiada, una entrevista de diagnóstico o terapéutica puede proporcionar una gran cantidad de información acerca de una persona: la naturaleza, duración y gravedad de sus problemas; cómo se manifiestan los problemas (hacia el interior o hacia el exterior); qué influencias pasadas están relacionadas con las dificultades presentes; los recursos y limitaciones del entrevistado para afrontar los problemas; los tipos de ayuda psicológica que el entrevistado ha recibido en el pasado, y los tipos de ayuda que se esperan y podrían ser de utilidad actualmente.

Método clínico e investigación de la moralidad

Sigmund Freud, Jean Piaget y otros psicólogos famosos usaron con mucha frecuencia el método clínico de entrevista, en el cual el entrevistador formula preguntas de sondeo para probar los límites u obtener información a profundidad acerca de una persona. El uso de la entrevista clínica en la investigación, conocido como *método clínico*, requiere de habilidad considerable.

TABLA 15.2 Recomendaciones para conducir una entrevista clínica

1. Asegure al entrevistado la confidencialidad de la entrevista.
2. Transmita un sentimiento de interés y calidez (*rapport*).
3. Trate de que el entrevistado se sienta cómodo.
4. Trate de entrar en contacto con los sentimientos del entrevistado (empatía).
5. Sea cortés, paciente y muestre aceptación.
6. Anime al entrevistado a expresar con libertad sus pensamientos y sentimientos.
7. Ajuste las preguntas a los antecedentes culturales y educativos del entrevistado.
8. Evite la jerga psiquiátrica o psicológica.
9. Evite las preguntas orientadoras.
10. Comparta información y experiencias personales con el entrevistado (autorrevelación) si resulta apropiado y el tiempo lo permite.
11. Utilice el humor con moderación y sólo si es apropiado y no ofensivo.
12. Escuche sin mostrar una reacción emocional excesiva.
13. Atienda no sólo a lo que se dice, sino también a cómo se dice.
14. Tome notas o haga un registro de la manera menos notoria posible.

Ejemplo de un instrumento de investigación que involucra el uso del método clínico es la Escala de Juicio Moral de Lawrence Kohlberg. Kohlberg (1969, 1974) sostenía que el desarrollo de la moralidad personal progresa a través de tres niveles ascendentes, cada uno de los cuales consta de dos etapas. En el nivel más bajo (*nivel premoral*), los juicios morales son guiados por el castigo y la obediencia o por una especie de filosofía ingenua de placer-dolor. En un nivel intermedio (*moralidad de conformidad con las reglas convencionales*), la moralidad depende de la aprobación de otras personas (la moralidad de "niño bueno o niña buena") o de la adherencia a los preceptos de la autoridad. En la primera etapa del último nivel (*moralidad de los principios morales aceptados por la persona*), la moralidad es vista en términos de la aceptación de un contrato o acuerdo determinado de manera democrática. En la segunda etapa del último nivel, el individuo ha desarrollado un conjunto interno de principios y una conciencia que dirige su juicio y su comportamiento.

La Escala de Juicio Moral se aplica presentando nueve dilemas morales hipotéticos y obteniendo los juicios del examinado y sus razones para emitir los juicios correspondientes a cada dilema. Uno de esos dilemas, el caso de Heinz y el farmacéutico, es el siguiente:

> En Europa, una mujer estaba próxima a morir de un tipo especial de cáncer. Existía un medicamento que los doctores pensaban podría salvarla. Se trataba de una forma de radio que un farmacéutico de la misma ciudad había descubierto recientemente. La elaboración del medicamento era costosa, pero el farmacéutico estaba cobrando diez veces más de lo que le costaba elaborarlo. Él había pagado $200 por el radio y cobraba $2,000 por una pequeña dosis del medicamento. El marido de la mujer enferma, Heinz, acudió a todo el que conocía para obtener prestado el dinero, pero sólo pudo reunir alrededor de $1,000, lo cual era la mitad del costo. Le dijo al farmacéutico que su esposa estaba muriendo y le pidió que se lo vendiera más barato o le permitiera pagárselo después. Pero el farmacéutico dijo, "No, yo descubrí el medicamento y voy a hacer dinero con él". Por lo que Heinz se desesperó e irrumpió en la tienda del hombre para robar el medicamento para su esposa. (Kohlberg y Elfenbein, 1975).

La calificación de los juicios morales del examinado y de las razones para emitir los juicios concernientes a historias como ésta consiste en hacer evaluaciones más bien subjetivas de esas respuestas en términos de las etapas de Kohlberg. Además de la subjetividad de la calificación, el enfoque de Kohlberg hacia el desarrollo moral ha recibido otras críticas. Una revisión de la evidencia concerniente a este enfoque hizo notar varios problemas conceptuales y metodológicos, incluyendo problemas en la derivación, administración y calificación de la Escala de Juicio Moral (Kurtines y Greif, 1994).

Entrevista de estrés

La regla habitual de cordialidad hacia el entrevistado se suspende en una *entrevista de estrés*. La meta de la entrevista de estrés, la cual se emplea en contextos clínicos, de selección y de interrogatorios policíacos, es determinar la habilidad de la persona para afrontar o resolver un problema específico bajo condiciones emocionales tensionantes. La entrevista de estrés también puede ser apropiada cuando no se dispone de mucho tiempo o cuando el entrevistado es muy repetitivo, indiferente o está a la defensiva. Se hace un intento por producir una respuesta emocional válida —ver por debajo de la máscara social superficial (el *personaje*) del entrevistado— formulando preguntas de sondeo y desafiantes en una atmósfera similar a un interrogatorio policiaco. Es obvio que se requiere mucha experiencia profesional para hacer que una entrevista de estrés parezca realista y evitar que las reacciones se salgan de control.

Entrevista cognoscitiva

La información de los testigos presenciales es claramente importante en la investigación de un delito, pero por lo general la policía recibe un entrenamiento inadecuado para entrevistar a testigos cooperativos. Un procedimiento de entrevista que en tiempos recientes se ha enseñado a muchos entrevistadores policiacos que conducen interrogatorios es la *entrevista cognoscitiva*. Este procedimiento fue desarrollado por Geiselman, Fisher, MacKinnon y Holland (1985) para obtener más información detallada y precisa de testigos de actos delictivos. La versión original del procedimiento de entrevista cognoscitiva consta de (1) inducir al testigo a recrear mentalmente el contexto original del delito, (2) dar instrucciones al testigo para que informe todo lo que observó, (3) lograr que el testigo recuerde los rasgos y acontecimientos de la escena del delito en distintos órdenes y (4) hacer que el testigo describa el suceso desde una variedad de perspectivas. Este procedimiento básico ha sido mejorado para abordar la dinámica social y la comunicación entre el entrevistador y el testigo (Fisher, Geiselman, Raymond, Jurkevich y Warhaftig, 1987; Fisher, McCauley y Geiselman, 1994). Los hallazgos de la investigación indican que las entrevistas cognoscitivas producen detalles más correctos que las entrevistas estructuradas (Köhnken, Schimossek, Aschermann y Höfer, 1995; Mantwill, Koehnken y Aschermann, 1995; McCauley y Fisher, 1995).

Entrevistas de personal

Casi todas las organizaciones de producción y servicios utilizan entrevistas, no sólo para la selección, clasificación y ubicación de los empleados, sino también para asesoramiento, resolución de problemas, liquidación (entrevista de salida) e investigación. Debido a que el proceso de entrevista es costoso y se lleva mucho tiempo, es razonable preguntarse si es el procedimiento más eficiente para obtener datos de los solicitantes de empleo. Buena parte de la información obtenida de una entrevista estructurada, la cual es el enfoque preferido en la mayor parte de los escenarios laborales, puede obtenerse de un formulario de solicitud o de un cuestionario. Sin embargo, los solicitantes de empleo con frecuencia se muestran más dispuestos a revelar asuntos de importancia en la atmósfera personal de una entrevista que por escrito. En la mayoría de los escenarios organizacionales, para todos los puestos salvo para los de menor nivel, una entrevista de personal es el paso final en el proceso de selección de empleados.

Los entrevistadores de personal disponen, por lo general, de información diversa acerca del solicitante, incluyendo la proporcionada por la forma de solicitud, las cartas de recomendación, calificaciones de pruebas y fuentes similares. La tarea del entrevistador es integrar la información obtenida de todas esas fuentes y la entrevista de personal para emitir una recomendación o tomar una decisión de empleo.

Un entrevistador de personal debe mostrarse cauto al hacer preguntas relativas a asuntos privados, no sólo porque pueden colocar al entrevistado bajo presión emocional, sino también porque puede ser ilegal plantearlas. En el cuadro 15.1 se presentan ejemplos de preguntas que son permisibles y otras que no lo son.

Confiabilidad y validez de las entrevistas

La entrevista es una herramienta psicológica importante, pero comparte con los métodos de observación los problemas de confiabilidad y validez. La confiabilidad requiere consistencia, pero los entrevistadores varían en su apariencia, enfoque, estilo y, en consecuencia, en la impresión

CUADRO 15.1
PREGUNTAS DE EMPLEO PERMITIDAS Y NO PERMITIDAS

Los lineamientos interpretativos publicados por la Comisión de Oportunidades Laborales Equitativas (estadounidense) indican que es permisible preguntar lo siguiente en una entrevista de trabajo:

¿Cuántos años de experiencia tiene?
(A una ama de casa) ¿Por qué desea volver a trabajar?
¿Cuáles son sus metas profesionales?
¿Quiénes han sido sus patrones anteriores?
¿Por qué dejó su empleo anterior?
¿Es usted un veterano de guerra? ¿El servicio militar le proporcionó alguna experiencia relacionada con el trabajo?
Si no tiene teléfono, ¿cómo podemos localizarlo?
¿Qué idiomas habla con fluidez?
¿Puede viajar?
¿Quién lo recomendó con nosotros?
¿Qué le gustó o le disgustó de sus empleos anteriores?
¿Cuáles son sus antecedentes educativos? ¿A qué escuelas asistió?
¿Cuáles son sus puntos fuertes? ¿Sus debilidades?
¿Tiene algún inconveniente en que verifiquemos sus referencias con su patrón anterior?

Por otro lado, se considera legalmente inaceptable preguntar lo siguiente en una entrevista de trabajo:

¿Qué edad tiene?
¿Cuál es su fecha de nacimiento?
¿Tiene hijos? De ser así, ¿cuántos años tienen?
¿Cuál es su raza?
¿A qué iglesia asiste?
¿Es usted casado, divorciado, separado, viudo o soltero?
¿Alguna vez ha sido arrestado?
¿Qué tipo de licencia militar tiene?
¿A qué clubes u organizaciones pertenece?
¿Su casa es rentada o propia?
¿A qué se dedica su esposa (esposo)?
¿Quién vive en su casa?
¿Alguna vez le han incautado o embargado sus bienes?
¿Cuál era su nombre de soltera (solicitantes mujeres)?

que causan en los entrevistados. Las impresiones diferentes producen diferencias en la conducta: una persona puede ser amistosa y comunicativa con un entrevistador, mientras que con otro se torna hostil y distante. Además, las percepciones que el entrevistador tiene del entrevistado pueden ser distorsionadas por sus experiencias y personalidad.

La confiabilidad de una entrevista se determina, por lo regular, comparando las calificaciones dadas a las respuestas del entrevistado por dos o más jueces. La magnitud de un coefi-

ciente de confiabilidad entre calificadores calculado a partir de esas calificaciones varía con la especificidad de las preguntas planteadas y las conductas calificadas; por lo general, es más alta para las entrevistas estructuradas y semiestructuradas que para las no estructuradas (Borman, Hanson y Hedge, 1997; Campion, Pursell y Brown, 1988). Sin embargo, aun cuando las preguntas sean bastante objetivas y se planteen en un formato estructurado, la confiabilidad entre calificadores de los datos de la entrevista usualmente no es mayor de .80.

Cuando se conduce una entrevista, el entrevistador es el instrumento de evaluación. En consecuencia, muchos de los problemas de confiabilidad de las entrevistas se relacionan con las características y conducta del entrevistador. Debido a que éste casi siempre está a cargo de la situación de entrevista, su personalidad y sus sesgos son, por lo regular, más importantes que los del entrevistado en la determinación del tipo de información obtenida. El tono socioemocional de una entrevista está determinado más por las acciones del entrevistador que por las del entrevistado: el entrevistador habla más y la extensión de las respuestas del entrevistado está directamente relacionada con la extensión de las preguntas formuladas por el entrevistador. Además de ser abiertamente dominante, el entrevistador puede no lograr obtener información completa y precisa al hacer preguntas erróneas, al no alentar respuestas completas o al no conceder tiempo suficiente para las mismas, y al registrar las respuestas de manera incorrecta.

Otros defectos de los entrevistadores son la tendencia a dar más peso a la primera impresión y a ser más afectados por la información desfavorable que por la favorable concerniente a un entrevistado. Los errores que afectan las calificaciones también ocurren en los juicios del entrevistador. Un ejemplo es el *efecto de halo*, que consiste en emitir juicios consistentemente favorables o desfavorables sobre la base de una "impresión general" o de una sola característica destacada del entrevistado. Esto ocurre cuando una persona que en realidad es superior (o inferior) en sólo una o dos características recibe una evaluación general superior (o inferior). También puede ocurrir un *error de contraste*, el cual consiste en juzgar a un entrevistado promedio como inferior si el entrevistado precedente fue claramente superior, o como superior si el entrevistado anterior fue claramente inferior.

Debido a que las impresiones del entrevistador son influidas por la limpieza, postura y otras conductas no verbales del entrevistado, así como por sus respuestas verbales, los futuros entrevistados harían bien en prepararse en lo mental y lo físico para una entrevista. En el caso de una entrevista laboral, el entrevistado deberá tener algún conocimiento de la organización y su filosofía. Deberá estar preparado para proporcionar una sinopsis de sus antecedentes y aspiraciones, pero abstenerse de hacer comentarios controvertidos o exhibir malos hábitos como fumar o morderse las uñas durante la entrevista (vea el cuadro 15.2).

Un hallazgo consistente de viejos estudios que atañen a la validez de la entrevista en la selección laboral o el diagnóstico clínico es que ésta se sobrestima (Arvey, 1979; Reilly y Chao, 1982). Revisiones más recientes (Borman, Hanson y Hedge, 1997; Maurer y Fay, 1988) subrayan el hecho de que las entrevistas pueden hacerse más válidas mediante la planeación y estructuración cuidadosas y el entrenamiento minucioso de los entrevistadores. Los resultados de una entrevista tienen mayor validez cuando el entrevistador (de preferencia más de uno) se centra en información específica (de trabajo o clínica) y las respuestas se evalúan pregunta a pregunta (de preferencia por dos o más evaluadores), más que como un todo. Para facilitar este proceso, toda la entrevista debe registrarse electrónicamente para su reproducción y evaluación posterior. Así, la tarea de interpretar las respuestas de un entrevistado puede separarse de manera más efectiva del proceso real de la entrevista. Pero no es suficiente con el registro de una entrevista en una cinta de vídeo, y especialmente en una cinta de audio. Las palabras habladas y las imágenes no siempre son claras, y el tono emocional y las variables contextuales con frecuencia se pierden en un

CUADRO 15.2

LO QUE NO SE DEBE HACER EN UNA ENTREVISTA LABORAL Y DESATINOS EN EL CURRÍCULUM*

LO MÁS IMPORTANTE QUE "NO DEBE HACERSE" EN LAS ENTREVISTAS

No pregunte "¿Cuánto tiempo va a durar?"
No diga "Soy una persona sociable".
No diga "Dejé los tres últimos puestos porque mi jefe se metía conmigo".
No pregunte "¿Cuánto tiempo de vacaciones voy a tener?"
No diga "No estoy seguro de lo que quiero hacer".
No pregunte "¿Puede firmar mi tarjeta de desempleo?"
No lleve un vestido de fiesta azul metálico.
No lleve pantalones cortos.
No deje expuesto su tatuaje.
No se quede dormido.
No lleve a sus hijos.
No lleve un refresco.
No lleve su radiolocalizador.

ALGUNOS DESATINOS FAVORITOS EN EL CURRÍCULUM

"Mi objeción profesional es..."
"Experiencia en relaciones privadas..."
"Con habilidades en corrección de prole..."
"Deseo trabajar para una compañía en la cual pueda ser menospreciado".
"Tengo el propósito de hablar y habilidades de langosta".
"Quiero un puesto para pagar mis cuentas".

*Recopilado por Servicios de Personal Snelling.

registro electrónico. Por esta razón, se necesita un observador humano alerta que tome buenas notas para complementar el registro electrónico de una entrevista.

Entrevista por computadora

A menudo la entrevista psicodiagnóstica puede automatizarse almacenando en una computadora un conjunto de preguntas e instrucciones. La computadora pregunta, recibe una respuesta y decide (*ramificación condicional*) qué pregunta debe ir a continuación. La estrategia de ramificación ha sido aplicada de manera eficaz a los sistemas de datos de los pacientes en muchos hospitales psiquiátricos.

En los años recientes ha aumentado el uso de la entrevista computarizada con propósitos de obtener historias de caso, conducir evaluaciones del comportamiento, concentrarse en problemas específicos, identificar síntomas-objetivo y ayudar en el diagnóstico psiquiátrico. Un paquete de software de computadora para la entrevista psicodiagnóstica y la preparación del informe correspondiente es el llamado Entrevista de Diagnóstico para Niños y Adolescentes IV (DICA-IV) Programa de Computadora para Windows (de W. Reich, Z. Weiner y B. Herjanic; Multi-Health Systems). La entrevista telefónica asistida por computadora (CATI) puede reali-

zarse con instrumentos como la Evaluación de Trastornos Mentales de Atención Básica (Kobak *et al*., 1997), los Exámenes del Estado Actual (Dignon, 1996), y el Programa de Entrevistas de Diagnóstico (Alhberg, Tuck y Allgulander, 1996; Bucholz, Marion, Shayka, Marcus y Robins, 1996). También se dispone de "computadoras parlantes" que conducen entrevistas sobre temas delicados, en particular en los casos de abuso infantil (por ejemplo, Romer *et al*., 1997).

Como sucede con otras aplicaciones psicométricas de las computadoras, las ventajas de la entrevista computarizada son eficiencia, flexibilidad y confiabilidad. La entrevista basada en la computadora ahorra tiempo profesional, permite la cobertura más amplia de temas y es más flexible que una serie de preguntas planteadas por un entrevistador humano. En general, existe un alto grado de acuerdo entre la información obtenida en la entrevista por computadora y la recopilada mediante entrevistas y cuestionarios psiquiátricos estándar. La mayoría de las personas no objetan ser entrevistadas mediante una computadora y, de hecho, pueden tener mayor disposición a divulgar información personal, en particular de naturaleza delicada, a una computadora impersonal que no emite juicios que a un entrevistador humano (Farrell, 1993; Feigelson y Dwight, 2000; Supple, Aquilino y Wright, 1999).

Entre las desventajas de la entrevista basada en computadora se encuentran que puede ser necesario abreviar o desviar el sistema en casos de crisis, que tiene utilidad limitada con niños y adultos de baja mentalidad, y que puede no ser lo suficientemente flexible como para usarla con una amplia gama de problemas y síntomas encontrados en los pacientes psiquiátricos. Otras desventajas potenciales de la entrevista basada en computadora incluyen dificultades para manejar otra cosa que no sea información verbal estructurada y la incapacidad para adaptar el planteamiento de las preguntas a la persona y el contexto. Una entrevista secuencial no estructurada, en la cual las preguntas sucesivas son determinadas por las respuestas del entrevistado a las preguntas previas, es más difícil de programar que un procedimiento de entrevista estructurada, en el cual se plantean las mismas preguntas a cada entrevistado.

EVALUACIÓN Y ANÁLISIS DEL COMPORTAMIENTO

El término *modificación del comportamiento* se refiere a un conjunto de procedimientos psicoterapéuticos basados en la teoría e investigación del aprendizaje y diseñados para cambiar la conducta inapropiada por un comportamiento personal y/o socialmente más adecuado. Las conductas inapropiadas pueden ser excesos, déficit u otras inadecuaciones de la acción susceptibles de ser corregidas mediante técnicas conductuales como la desensibilización sistemática, el contracondicionamiento y la extinción. Entre las conductas inadaptadas que han recibido atención especial de los modificadores del comportamiento se encuentran temores específicos (o fobias), tabaquismo, comer en exceso, alcoholismo, adicción a drogas, falta de asertividad, mojar la cama, tensión y dolores crónicos, y problemas sexuales. Aunque esas conductas meta por lo regular se definen de manera limitada, los terapeutas conductuales de orientación más cognoscitiva también han abordado problemas más generales, como el autoconcepto negativo y la crisis de identidad. Además, las conductas meta no sólo constan de movimientos no verbales, sino también de informes verbales de pensamientos y sentimientos.

Análisis del comportamiento

Los terapeutas conductuales intentan entender la conducta mediante la identificación de sus antecedentes, lo cual incluye tanto el historial de aprendizaje social como el entorno presente y los

resultados o consecuencias de esta conducta. Un principio fundamental de la modificación del comportamiento, basado en los estudios de laboratorio del aprendizaje operante, es que la conducta está controlada por sus consecuencias. Al diseñar un programa para corregir una conducta problema, debemos identificar no sólo las condiciones que la preceden y desencadenan, sino también las consecuencias reforzantes que la mantienen. Al usar este enfoque, el proceso de modificación del comportamiento es precedido por un *análisis funcional* de la(s) conducta(s) problema. Este análisis consiste en una secuencia A-B-C en la cual A representa las condiciones antecedentes, B la conducta problema y C las consecuencias de esa conducta. B se modifica controlando A y alterando C. Los antecedentes y las consecuencias de la conducta meta pueden ser condiciones manifiestas observables de manera objetiva o eventos mentales encubiertos reportados por la persona cuya conducta es modificada.

Evaluación conductual

La evaluación conductual tiene funciones múltiples, incluyendo (1) la identificación de las conductas meta, conductas alternativas y variables causales; (2) el diseño de las estrategias de intervención, y (3) la reevaluación de las conductas meta y causal (Haynes, 1990). Se emplean varios procedimientos, incluyendo observaciones y entrevistas, además de listas de verificación, escalas de calificación y cuestionarios completados por el paciente o por una persona familiarizada con él. En ocasiones, los modificadores del comportamiento han usado incluso las respuestas a técnicas proyectivas como muestras de conducta (vea Maloney y Ward, 1976).

Métodos observacionales. Los procedimientos observacionales empleados en un análisis del comportamiento implican tomar nota de la frecuencia y duración de las conductas meta y las contingencias particulares (antecedentes y consecuencias) de su ocurrencia. Dependiendo del contexto y de la edad del paciente, las observaciones conductuales pueden ser hechas y registradas por maestros, padres, enfermeras, asistentes de enfermería o por cualquier otra persona que esté familiarizada con el paciente.

Autosupervisión. La autoobservación puede ser la forma más sencilla y económica de determinar con qué frecuencia y bajo qué condiciones ocurre una conducta meta particular. Aunque la autoobservación no siempre es confiable, la gente puede ser entrenada para efectuar observaciones precisas y válidas de su propia conducta (Kendall y Norton-Ford, 1982). En la autoobservación con propósitos de análisis y modificación del comportamiento, se indica a la persona portar todo el tiempo materiales como una libreta, un contador de pulsera y un cronómetro para llevar un registro de las ocurrencias de la conducta meta y del momento, lugar y circunstancias en que ocurra. La autoobservación, o *autosupervisión*, puede ser bastante confiable cuando se entrena con cuidado al paciente. Es interesante que el mismo proceso de autosupervisión —observar y tabular las ocurrencias de conductas específicas en que se ocupa un individuo— pueda afectar la ocurrencia de esas conductas, a menudo de manera terapéutica (Ciminero, Nelson y Lipinski, 1977). Por ejemplo, las personas que fuman mucho tienden a hacerlo menos cuando llevan un registro de qué tan a menudo, por cuánto tiempo y en qué circunstancias fuman. Al tomar mayor conciencia del tabaquismo, éste se vuelve menos automático y se pone bajo un mayor control consciente.

Entrevista conductual. La entrevista conductual es un tipo de entrevista clínica concentrado en obtener información para planear un programa de modificación del comportamiento. Esto

implica presentar al entrevistado una descripción objetiva de la conducta problema, así como de las condiciones antecedentes y las consecuencias reforzantes. La conducción exitosa de una entrevista conductual requiere alentar y enseñar al entrevistado a responder en términos de conductas específicas, más que en el lenguaje usual de motivos y rasgos. Después de obtener la información necesaria para desarrollar un programa de modificación del comportamiento, se le explica a la persona y ésta debe estar motivada para perseverar con el programa.

RESUMEN

Observaciones y entrevistas son los métodos de mayor uso, pero no necesariamente los más válidos, para evaluar la personalidad. Las observaciones pueden ser controladas o no controladas y formales o informales. Otros tipos de observación son naturalista, participante y las llamadas autoobservaciones. Las observaciones naturalistas ocurren en situaciones naturales más que en situaciones arregladas de antemano. Las observaciones participantes se hacen cuando el observador se vuelve un participante en el grupo que está siendo observado. Los documentos personales que resultan de las autoobservaciones son evaluados por medio del análisis de contenido.

En una prueba de situación se observa a un participante auténtico para determinar con cuánta efectividad puede resolver un problema asignado bajo circunstancias frustrantes. A pesar de su realismo, las pruebas de situación no son tan válidas como puede parecer al principio.

La confiabilidad y la validez de las observaciones objetivas pueden ser mejoradas mediante el muestreo de tiempo e incidentes, el entrenamiento cuidadoso de los observadores, conducir la entrevista de manera tan discreta como sea posible, y el registro electrónico. Debe entrenarse a los observadores para atender a conductas verbales y no verbales.

Información sobre la historia de vida de una persona puede obtenerse de manera eficiente a partir de un formulario de solicitud o un inventario biográfico, además de las conversaciones con gente que conozca a la persona. Las cartas de recomendación también se usan de manera extensiva, pero a menudo son de valor cuestionable. Esto es cierto, sobre todo cuando la persona que hace la recomendación sabe que la carta puede ser leída por la persona sobre la que se escribió.

Dependiendo de sus propósitos y de las habilidades de los entrevistadores, las entrevistas pueden ser estructuradas, semiestructuradas o no estructuradas. Pueden ser conducidas con propósitos clínicos, educativos, de empleo, entre otros. Las entrevistas del estado mental se realizan con propósitos legales para determinar la competencia mental. Las entrevistas de estrés involucran el uso de un enfoque de confrontación diseñado para romper la resistencia y las defensas. La entrevista cognoscitiva fue diseñada principalmente para obtener información detallada y más precisa en situaciones de interrogatorio policiaco.

Se ha publicado una serie de programas de entrevista estándar, sobre todo para uso en situaciones clínicas. Además de la entrevista tradicional cara a cara, algunas entrevistas son conducidas por computadora y/o por teléfono. La confiabilidad de las entrevistas es bastante modesta, pero puede mejorarse incrementando su estructura, el entrenamiento minucioso de los entrevistadores y puntualizando un registro meta de los resultados de la entrevista.

Los procedimientos especiales de observación incluyen los registros anecdóticos, el muestreo de tiempo, el muestreo de incidentes y la prueba de situación. Un tipo especial de entrevista es la entrevista de estrés, un enfoque de confrontación que requiere de mucho entrenamiento para ser efectivo. Las entrevistas se realizan con varios propósitos, pero ciertos tipos de preguntas de entrevista se consideran ilegales en la selección de personal.

Tanto la observación como la entrevista se utilizan en el análisis del comportamiento y en el diseño de programas de modificación de conducta. El análisis del comportamiento consiste en la aplicación de varias técnicas para obtener información acerca de un paciente cuya conducta es inadaptada de alguna manera. Un análisis del comportamiento resulta en la especificación de las condiciones antecedentes (A), las conductas inadaptadas meta (B), y las consecuencias de esas conductas (C). El proceso de modificación del comportamiento consiste en arreglar la situación de forma que la conducta meta no sea desencadenada por ciertos estímulos o seguida por determinadas consecuencias.

PREGUNTAS Y ACTIVIDADES

1. Seleccione a alguno de sus compañeros de clase como sujeto de observación, de preferencia alguien a quien usted no conozca y hacia quien tenga sentimientos neutrales. Observe a la persona por un periodo de tres o cuatro clases y, de manera discreta, registre lo que hace y lo que dice. Trate de ser tan objetivo como sea posible, busque conductas consistentes y típicas, y tome nota de las respuestas que ocurren con poca frecuencia. Al final del periodo de observación escriba una caracterización de dos a tres páginas de la persona. Sin tener acceso a cualquier otra información acerca de la persona (lo que otros estudiantes dicen acerca de ella, lo bien que le va en la escuela, y cosas similares), ¿cómo describiría su personalidad y conducta característica? Por último, verifique sus observaciones con las de otros individuos que conozcan o hayan observado a la persona de su estudio. Después de esta experiencia de observación cercana usando una técnica de muestreo de tiempo, ¿cómo se siente acerca de la observación objetiva como método de evaluación de la personalidad? ¿Es confiable, válido y útil?

2. Pida a seis personas, una a la vez, que hagan expresiones faciales indicativas de cada una de las siguientes emociones: ira, repugnancia, temor, felicidad, tristeza y sorpresa. Tome notas sobre las expresiones faciales, diferenciando entre las diversas emociones. ¿Encontraron sus "actores" difícil la tarea? ¿Hubo consistencia apreciable de una persona a otra en las expresiones que caracterizan a una emoción en particular? ¿Fueron ciertas emociones más fáciles de expresar y se expresaron con más consistencia que otras?

3. Revise el análisis de los procedimientos de entrevista tratados en este capítulo y otros lineamientos de entrevista a su alcance. Conduzca luego una entrevista personal estructurada de alguien a quien no conozca bien. Escriba los resultados como un informe formal en el que dé la información de identificación, un resumen de los hallazgos de la entrevista y recomendaciones concernientes al entrevistado.

4. Prepare una lista de preguntas a ser planteadas durante una entrevista laboral y conduzca la entrevista con un conocido. Siéntase en libertad de desviarse del programa de la entrevista si piensa en preguntas que sean más pertinentes para el desempeño de la persona (solicitante) en el(los) trabajo(s) que solicita. Asegúrese de que todas las preguntas planteadas estén relacionadas con el trabajo y que sean legalmente permisibles.

5. ¿Cuáles son algunas ventajas y desventajas de los procedimientos de observación y entrevista en la evaluación de la personalidad? Concéntrese en la simplicidad, objetividad, confiabilidad y validez relativas, las situaciones de evaluación para las cuales son apropiados, y cualquier otra ventaja o desventaja que posean los procedimientos de observación y entrevista.

6. ¿Cuáles son las diferencias entre las entrevistas planeadas y no planeadas, controladas y no controladas, estructuradas y no estructuradas?
7. ¿En qué situaciones o circunstancias sería apropiado usar la observación participante? ¿Qué tipos de información puede esperarse proporcione la observación participante y cuáles son sus defectos?
8. ¿Qué es el análisis del comportamiento y qué papel juegan en este proceso la observación y las entrevistas?
9. ¿Cuáles son algunos de los factores a tomar en consideración al entrenar a los entrevistadores? ¿En qué medida los buenos entrevistadores nacen más que se hacen?

CAPÍTULO DIECISÉIS

LISTAS DE VERIFICACIÓN Y ESCALAS DE CALIFICACIÓN

La información obtenida a partir de observaciones y entrevistas, de manera formal o informal, puede registrarse de diversos modos. Debido a la enorme masa de datos producida en las largas sesiones de observación y entrevista, los resultados casi siempre se resumen de algún modo. Junto con una descripción condensada por escrito, las listas de verificación y las escalas de calificación son instrumentos útiles para resumir los datos obtenidos a partir de observaciones y entrevistas. Si bien los reactivos de las listas de verificación, por lo regular, sólo requieren respuestas dicotómicas (presente/ausente, sí/no y así por el estilo), en algunas listas de verificación se proporcionan tres opciones (marca sí, marca no, o sin marca). En las escalas de calificación se pide a la persona que responde formular juicios evaluativos sobre una serie ordenada de tres o más categorías.

Superadas en popularidad sólo por las pruebas de aprovechamiento, las listas de verificación y las escalas de calificación son instrumentos psicométricos convenientes, económicos y versátiles. Pueden ser elaboradas con facilidad, aplicadas de manera conveniente con sólo lápiz y papel, utilizadas para describirse uno mismo, describir a alguien o algo más, y ser adaptadas a la medición de una amplia gama de conductas, características personales y otros objetos, acontecimientos o condiciones. En el mercado pueden encontrarse cientos de listas de verificación y escalas de calificación. Esos instrumentos pueden administrarse solos o en combinación con otros métodos para evaluar a la gente y con otros propósitos.

CARACTERÍSTICAS DE LAS LISTAS DE VERIFICACIÓN

Una *lista de verificación* es un método relativamente sencillo, económico y bastante confiable para describir o evaluar a una persona. Consiste en una lista de palabras, frases o afirmaciones descriptivas de una persona o algún objeto o acontecimiento. Elaboradas con mayor facilidad que las escalas de calificación o los inventarios de personalidad, y a menudo de igual validez, las listas de verificación pueden aplicarse como instrumento de autorreporte o de informe de un observador. Se pide a los examinados que marquen, subrayen o indiquen de alguna otra manera, qué palabra(s) o frase(s) los describe (autoverificación) o describe a alguien o algo más. Las listas de verificación son más eficientes porque, a diferencia de las escalas de calificación, no requieren que el individuo tome decisiones explícitas acerca de la calidad, frecuencia o intensidad de las conductas y características. Las escalas de calificación pueden proporcionar información más detallada que las listas de verificación, pero se requiere más tiempo para completarlas. En consecuencia, es posible que exista una especie de trueque rapidez-exactitud entre los dos instrumentos.

Las listas de verificación se usan con gran frecuencia en los contextos clínico, educativo e industrial-organizacional. Aunque algunas listas de verificación son instrumentos estandariza-

dos y están disponibles de manera comercial, muchas han sido preparadas con propósitos especiales o para usarse en contextos específicos. Por ejemplo, la lista de verificación del formato 16.1 es un instrumento no estandarizado, diseñado para medir la conducta tipo A. Otros dos ejemplos de listas de verificación no estandarizadas son la Escala de Calificación de Readaptación Social (Holmes y Rahe, 1967) y la Lista de Verificación Conductual para la Ansiedad en el Desempeño (Paul, 1966).

Escala de Reajuste Social

La Escala de Reajuste Social (SRS) fue diseñada para estudiar los efectos de los cambios de la vida, negativos y positivos, en la conducta y las reacciones fisiológicas al estrés producido por esos cambios (Holmes y Rahe, 1967). La teoría en la que se basa la SRS de 43 reactivos asume que entre mayor sea el grado de reajuste en un año dado, mayor es la probabilidad de que la persona desarrolle una enfermedad relacionada con el estrés. Cada reactivo en la SRS tiene un peso de calificación de 0 a 100, dependiendo del grado de reajuste requerido por el acontecimiento descrito en el reactivo. Después de evaluar las críticas concernientes a la SRS, Scully, Tosi y Banning (2000) concluyeron que es una herramienta útil para los investigadores y profesionales relacionados con el estrés.

Lista de Verificación Conductual para la Ansiedad en el Desempeño

En el formato 16.2 se muestra una segunda lista de verificación que no está disponible de manera comercial, la Lista de Verificación Conductual para la Ansiedad en el Desempeño. Este instrumento se utiliza para evaluar los efectos sobre la ansiedad de un tipo de terapia conductual conocido como *desensibilización sistemática*. Una ventaja de ésta y otras listas de verificación similares es que pueden llenarse de manera repetida o periódica para determinar si han ocurrido cambios en la conducta como resultado del tratamiento. Se marca en cada uno de los cuadros del formato 16.2 para indicar la ocurrencia de la conducta correspondiente durante el periodo designado (de 1 a 8).

Selección de una lista de verificación

Aunque las listas de verificación no estandarizadas como la anterior no por necesidad son provisionales o de mala calidad, rara vez son validadas de manera adecuada. En consecuencia, es

FORMATO 16.1 Lista de verificación descriptiva

Instrucciones: Coloque una marca de verificación en la línea para cada reactivo que lo describa.

___	**1.** orientado al logro	___	**11.** emocionalmente explosivo
___	**2.** agresivo	___	**12.** trabaja rápido
___	**3.** ambicioso	___	**13.** trabaja duro
___	**4.** competitivo	___	**14.** altamente motivado
___	**5.** trabajador constante	___	**15.** impaciente
___	**6.** le disgusta perder el tiempo	___	**16.** le agradan los desafíos
___	**7.** se molesta con facilidad	___	**17.** le agrada ser líder
___	**8.** se activa con facilidad	___	**18.** le agrada la responsabilidad
___	**9.** se frustra con facilidad	___	**19.** inquieto
___	**10.** eficiente	___	**20.** se esfuerza por tener éxito

FORMATO 16.2 Lista de Verificación Conductual para la Ansiedad en el Desempeño

	PERIODO							
CONDUCTA OBSERVADA	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1 Pasea								
2 Se balancea								
3 Arrastra los pies								
4 Tiemblan las rodillas								
5 Movimientos extraños del brazo y la mano (balanceos, rasguños, juguetear, etc.)								
6 Brazos rígidos								
7 Manos restringidas (en los bolsillos, detrás de la espalda, apretadas)								
8 Temblores en la mano								
9 No hay contacto ocular								
10 Músculos del rostro tensos (contraídos, tics, muecas)								
11 Rostro inexpresivo								
12 Rostro pálido								
13 Rostro sonrojado (rubor)								
14 Se humedece los labios								
15 Traga								
16 Aclara la garganta								
17 Respira con dificultad								
18 Suda (el rostro, las manos, las axilas)								
19 Voz temblorosa								
20 Se bloquea el habla o tartamudeo								

incierto si la lista de verificación está cumpliendo los propósitos para los que se creó. Por esta razón, es prudente considerar una de las listas de verificación disponibles de manera comercial antes de elaborar una nueva. En el mercado se dispone de listas de verificación de conducta adaptativa, progreso en el desarrollo, problemas de salud, características personales, historia personal, problemas personales y síntomas psicopatológicos. Se cuenta con listas de verificación para ansiedad, depresión, hostilidad, psicopatía y condición mental, así como con listas pertinentes para las relaciones matrimoniales, sexuales e interpersonales en los adultos (vea Aiken, 1996). Sean estandarizadas o no, estén disponibles o no en el mercado, al seleccionar cualquier lista de verificación o escala de calificación deben considerarse las siguientes preguntas:

1. ¿Qué variables (constructos) son medidas por el instrumento y cómo se definen?
2. ¿Cuál es la lógica sobre la que se basa el instrumento (una teoría específica de la personalidad o la conducta, resultados de investigaciones previas y temas similares)?
3. ¿Qué capacitación especial o condiciones específicas se requieren para usar el instrumento? ¿Por quién y bajo qué condiciones (contexto ambiental, materiales y aspectos similares) puede usarse?
4. ¿Cómo se califica el instrumento y qué materiales se necesitan para calificarlo? ¿Puede calificarse de manera rápida y precisa a mano o se necesita una computadora u otra máquina de calificación?
5. ¿Está estandarizado el instrumento? De ser así, ¿el grupo de estandarización era representativo de la gente que será evaluada con el instrumento?
6. ¿Qué tipos de evidencia se presentan para apoyar la confiabilidad (test-retest, formas paralelas, consistencia interna u otra) del instrumento?
7. ¿Qué tipos de evidencia se presentan o están disponibles en otras fuentes para apoyar la validez (de contenido, relacionada con el criterio, de constructo) del instrumento?

Calificación de la listas de verificación[1]

Una lista de verificación que consta de un conjunto de reactivos discretos, no relacionados, no se califica como una totalidad, sino que las respuestas a los reactivos individuales son examinadas, dentro y entre las personas que responden. Por supuesto, el número de personas que responden a un reactivo dado puede ser determinado y comparado con el de quienes responden a cada uno de los otros reactivos.

La calificación convencional de las respuestas a conjuntos interrelacionados de reactivos de listas de verificación designados para medir la misma variable por lo general empieza asignando un punto a cada reactivo marcado y cero puntos a cada reactivo no marcado; se da una calificación de +1 si la marca del reactivo indica una respuesta favorable, y una calificación de cero si la marca indica una respuesta desfavorable hacia cualquier cosa que pueda ser la variable expresada en el reactivo. En ciertos casos se asignan pesos de calificación distintos a cero y uno, como cuando los reactivos se escalan de acuerdo con su importancia. Sin embargo, cuando el número de reactivos es grande, dar diferentes pesos a diferentes reactivos por lo general tiene poco efecto sobre la confiabilidad o validez del instrumento. Cuando un número de individuos evalúa a la misma persona en una lista de verificación, es posible determinar una calificación de grupo en cada reactivo contando el número de individuos que lo marcaron.

Cuando a los examinados no se les indica marcar un cierto número de reactivos, distintos individuos pueden marcar un número diferente de reactivos. Debido a que este *conjunto frecuencia-respuesta*, como se le llama en ocasiones, puede tener un efecto pronunciado sobre las calificaciones globales, se necesita algún método de compensación. Por ejemplo, se proporcionan normas separadas en las diversas escalas de la Lista de Verificación de Adjetivos (ACL) para cada uno de los cinco grupos de intervalo de "Número Marcado". Para convertir la puntuación cruda de una persona en las escalas ACL a calificaciones estándar o normalizada, el calificador usa las tablas de conversión de puntuación cruda a calificación estándar normalizada presentada para el grupo en el cual el intervalo "Número Marcado" contiene el número de adjetivos marcados por la persona. Aiken (1996) describe otros métodos para controlar estadísticamente el conjunto frecuencia-respuesta.

[1]Es posible solicitar al autor un programa de cómputo para la elaboración y calificación de varios tipos de listas de verificación y escalas de calificación. Si está interesado en adquirirlo, envíe un sobre con estampillas y su dirección, y un diskete formateado en DOS, al doctor Lewis R. Aiken, 3300 Blue Ridge Court, Thousand Oaks, CA 91362.

Confiabilidad y validez

Las calificaciones (0 y 1) a los reactivos individuales en una lista de verificación tienen menor confiabilidad que las sumas de calificaciones de varios reactivos. Los coeficientes de confiabilidad para calificaciones sumadas entre reactivos pueden determinarse por medio de los métodos test-retest, consistencia interna y formas paralelas descritos en el capítulo 5. La confiabilidad de las listas de verificación determinada mediante esos procedimientos es, por lo regular, menor que la de las pruebas cognoscitivas. Un enfoque alternativo para determinar la confiabilidad de conjuntos de reactivos de listas de verificación es el método de acuerdo o concordancia entre verificadores. Este método consiste en calcular una sola calificación de acuerdo (phi) a partir de la concordancia de las configuraciones de marcas de verificación de dos o más verificadores (Sinacore, Connell, Olthoff, Friedman y Gecht, 1999).

Con respecto a la validez de las listas de verificación, los resultados de la investigación indican que las calificaciones de las listas de verificación tienen correlaciones significativas con una amplia gama de criterios de desempeño. Las calificaciones en las listas de verificación del desempeño de los empleados, la efectividad del tratamiento y otros criterios también tienen una relación significativa con las calificaciones en varias variables predictoras. Por ejemplo, Boyle y sus colaboradores (Boyle *et al*., 1996, 1997) encontraron que los coeficientes de confiabilidad y de validez de las listas de verificación de trastornos psiquiátricos eran similares, si no es que superiores, a los de las entrevistas. Por su parte, MacRae *et al*., (1995) encontraron que las calificaciones de las listas de verificación tendían a correlacionar más alto con las calificaciones de los médicos que las calificaciones de las bases de datos llenadas por estudiantes. En un estudio de las propiedades psicométricas de una lista de verificación estandarizada para el paciente y una forma de escala de calificación para evaluar las habilidades interpersonales y de comunicación, Cohen *et al*. (1996) encontraron que la confiabilidad de la forma de calificación era ligeramente más alta que la de la lista de verificación.

TIPOS Y EJEMPLOS DE LISTAS DE VERIFICACIÓN

Listas de verificación de adjetivos

Las listas de verificación que constan de una serie de adjetivos, como agresivo, ambicioso, competitivo, eficiente, explosivo, impaciente, irritable, inquieto y tenso, son muy populares y muy sencillas de elaborar. En ocasiones se dice que las personas descritas por esos nueve adjetivos tienen una personalidad tipo A (vea la sección de preguntas y ejercicios, punto 3). Dos de las listas de verificación de adjetivos estandarizadas de mayor popularidad son la Lista de Verificación de Adjetivos (ACL) (CPP) y la Lista de Verificación Múltiple de Adjetivos de Afecto (EdITS).

Lista de Verificación de Adjetivos (ACL). La Lista de Verificación de Adjetivos (ACL) consta de 300 adjetivos arreglados de manera alfabética desde *distraído* hasta *bromista* (*absent minded* to *zany*). A los examinados les lleva de 15 a 20 minutos marcar los adjetivos que consideran los describen. Esas respuestas pueden calificarse luego en las 37 escalas descritas en el manual de la ACL: 4 escalas de procedimientos, 15 de necesidades, 9 temáticas, 5 de análisis transaccional y 4 de originalidad-intelecto (creatividad e inteligencia). Las calificaciones en las escalas de procedimientos (número total de adjetivos marcados, número de adjetivos favorables marcados, número de adjetivos desfavorables marcados, comunales) atañen a la forma en que el sujeto manejó la lista de verificación. Las escalas de necesidades (escalas 5 a 19) están basadas en la descripción que hizo Edwards (1954) de las 15 necesidades en la teoría de la personalidad de necesidad-presión de Murray (1938). Cada una de las escalas temáticas (escalas 20 a

28) evalúa un tema o componente diferente de la conducta interpersonal (por ejemplo, preparación para la orientación, ajuste personal, personalidad creativa, atributos masculinos). Las escalas del análisis transaccional (escalas 29 a 33) se describen como medidas de las cinco funciones del yo en el análisis transaccional de Berne (1966). Las escalas de originalidad-intelecto (escalas 34 a 37) se describen como medidas de las dimensiones de personalidad de originalidad-intelecto (creatividad e inteligencia) de Welsh.

Para propósitos de interpretación y orientación, las puntuaciones crudas de la ACL se convierten a calificaciones *T* estándar. Como un ejemplo, en la tabla 16.1 se proporcionan las 37 calificaciones *T* y el perfil asociado de los casos descritos en el informe 16.1. Las calificaciones *T* se interpretan con referencia a normas basadas en muestras de 5,236 varones y 4,144 mujeres de 37 entidades estadounidenses. También se proporcionan los perfiles y las interpretaciones asociadas para seis casos de muestra, uno de los cuales se resume en el informe 16.1. La confiabilidad por consistencia interna de la mayoría de las 37 escalas es razonablemente alta, pero los datos de confiabilidad test-retest son limitados. El manual informa de coeficientes de confiabilidad test-retest para las escalas separadas que van desde .34 para la escala de alta originalidad, bajo inte-

INFORME 16.1 Descripción del caso que acompaña a las calificaciones de la Lista de Verificación de Adjetivos en la tabla 16.1

Esta estudiante universitaria de 19 años cursa una licenciatura en biología, ha mantenido un promedio de A y planea asistir a la escuela de posgrado. Creció en una familia numerosa y unida y tiene sentimientos cálidos hacia sus padres y su niñez. Antes de asistir a la universidad, siempre vivió en ciudades pequeñas o áreas semirurales. Asistir a una universidad de la ciudad requirió de gran adaptación, pero a ella le gustó la emoción y lo estimulante de la vida citadina. Mantuvo sus creencias religiosas y asistía a la iglesia con regularidad. Se considera conservadora en lo político y lo económico. La persona que la entrevistó acerca de la historia de su vida la describe de la siguiente manera:

> Es una joven inteligente, vivaz y atractiva, entusiasmada por su vida en la universidad. Aunque se ve a sí misma como introvertida, su conducta es más extrovertida, se mostró conversadora, sociable, sincera y no dudaba en asumir un papel de liderazgo. Sus padres fueron estrictos, esperaban que sus hijos asumieran responsabilidades y concedían gran valor al logro académico. Ella describió a su madre como una mujer exigente, muy tímida, que participaba en actividades sociales por un sentido de deber. Dijo que su padre era algo intimidante, pero afectuoso; se sentía más cercana a él ahora que cuando estaba creciendo. Estar en la escuela —lejos de casa y del relativo aislamiento de ese entorno— era muy emocionante.

Las calificaciones que obtuvo en el perfil de la ACL estaban de acuerdo con los datos de la historia de caso y con las evaluaciones del equipo. Se presentaron elevaciones moderadas en las escalas de Logro, Autoconfianza y Ajuste Personal y calificaciones de 60 o más en las escalas de Yo Ideal, Personalidad Creativa y A-2 (alta originalidad, alto intelecto). El perfil ACL también reveló calificaciones de 60 o más altas en las escalas para Favorable, Comunidad, Feminidad, Padre Crítico y A-4 (baja originalidad, alto intelecto). Aunque la calificación asignada por el equipo de 54 en Feminidad estaba por encima del promedio para la muestra de 80 estudiantes incluidos en este proyecto, no era tan alta como la calificación de 69 en la ACL autodescriptiva. Como obtuvo calificaciones superiores a 50 tanto en Masculinidad como en Feminidad, se le ubica en la casilla de andróginos en el diagrama de interacción entre las dos escalas. El perfil también revela calificaciones elevadas tanto en Favorable como en Desfavorable, lo cual sugiere que es más compleja, diferenciada en lo interior y menos represiva que sus compañeros.

TABLA 16.1 Escalas y calificaciones *T* de muestra en la Lista de Verificación de Adjetivos

NOMBRE Y DESIGNACIÓN DE LA ESCALA	CALIFICACIONES *T* PARA EL CASO DESCRITO EN EL INFORME 16.1	NOMBRE Y DESIGNACIÓN DE LA ESCALA	CALIFICACIONES *T* PARA EL CASO DESCRITO EN EL INFORME 16.1
Procedimientos		*Escala temática*	
1. Número total de adjetivos marcados (No Ckd)	37	**20.** Preparación para la orientación (Crs)	55
2. Número de adjetivos favorables marcados (Fav)	62	**21.** Autocontrol (S-Cn)	48
3. Número de adjetivos desfavorables marcados (Unfav)	59	**22.** Confianza en sí mismo (S-Cfd)	59
4. Comunales (Com)	68	**23.** Ajuste personal (P-Adj)	53
Escalas de necesidades		**24.** Yo ideal (Iss)	64
5. Logro (Ach)	57	**25.** Personalidad creativa (Cps)	63
6. Dominio (Dom)	50	**26.** Liderazgo militar (Mls)	52
7. Resistencia (End)	53	**27.** Atributos masculinos (Mas)	54
8. Orden (Ord)	57	**28.** Atributos femeninos (Fem)	69
9. Intracepción (Int)	57	*Análisis transaccional*	
10. Crianza (Nur)	44	**29.** Padre crítico (CP)	62
11. Afiliación (Aff)	53	**30.** Padre que cría (NP)	48
12. Heterosexualidad (Het)	46	**31.** Adulto (A)	56
13. Exhibición (Exh)	44	**32.** Niño libre (FC)	46
14. Autonomía (Aut)	49	**33.** Niño adaptado (AC)	41
15. Agresión (Agg)	58	*Originalidad-Intelecto*	
16. Cambio (Cha)	58	**34.** Alta originalidad, bajo intelecto (A-1)	47
17. Ayuda (Suc)	41	**35.** Alta originalidad, alto intelecto (A-2)	64
18. Humillación (Aba)	56	**36.** Baja originalidad, bajo intelecto (A-3)	44
19. Deferencia (Def)	49	**37.** Baja originalidad, alto intelecto (A-4)	63

lecto, hasta .77 para la escala de agresión (mediana de .65), y también describe muchos usos de la ACL e investigaciones en las que se ha utilizado.

Las revisiones de la ACL han sido bastante positivas y concluyen que el instrumento está bien desarrollado (Teeter, 1985; Zarske, 1985). Las escalas tienen una intercorrelación significativa y, por ende, no deberían interpretarse como factores independientes. Un análisis factorial que el autor de este libro realizó sobre las 15 escalas de necesidades (escalas 5 a 19) arrojó tres factores: autoconfianza o fortaleza del yo, orientación hacia la meta e interactividad social o amistad. La ACL se ha usado sobre todo con adolescentes y adultos normales, y no se ha determinado su validez en el psicodiagnóstico y la planeación del tratamiento. Se ha encontrado más útil en la investigación sobre el autoconcepto.

Lista de Verificación Múltiple de Adjetivos de Afecto, revisada. La Lista de Verificación Múltiple de Adjetivos de Afecto, revisada (MAACL-R) (Zuckerman y Lubin, 1985) consta de

132 adjetivos y se encuentra disponible para aplicarse de dos formas: rasgo ("En general") y estado ("Hoy"). Dependiendo de la forma, los examinados marcan aquellos adjetivos que indican cómo se sienten de manera general (en la forma para rasgos) o como se sienten el día del examen o en el presente (en la forma para estado). Se ha demostrado que ambas formas discriminan entre pacientes con trastornos afectivos y otros trastornos de los no pacientes. Se obtienen calificaciones *T* estándar en las formas de rasgo y de estado para cinco escalas básicas: Ansiedad (A), Depresión (D), Hostilidad (H), Afecto Positivo (PA) y Búsqueda de Sensaciones (SS). También pueden calcularse dos calificaciones estándar resumidas, Disforia (Dys = A + D + H) y Afecto Positivo y Búsqueda de Sensaciones (PASS = PA + SS). Las normas para la forma de rasgo de la MAACL-R se basan en una muestra nacional de 1,491 individuos de 18 años en adelante; las normas para la forma de estado se basan en una muestra (no representativa) de 538 estudiantes de una universidad del oeste medio. Con la excepción de la escala de Búsqueda de Sensaciones, los coeficientes de confiabilidad por consistencia interna para las escalas de rasgo y de estado son adecuados. La confiabilidad test-retest es satisfactoria para las escalas de rasgo, pero, como era de esperar de las fluctuaciones momentáneas en las actitudes y la conducta, es baja para las escalas de estado. En el manual de la MAACL-R (Zuckerman y Lubin, 1985) se presentan los resultados de los estudios de validez en varias poblaciones, incluyendo adolescentes y adultos normales, clientes que reciben orientación y pacientes de clínicas y hospitales estatales. Las calificaciones en la MAACL-R correlacionan en la dirección esperada con otras medidas de la personalidad (por ejemplo, el Inventario Multifásico de Personalidad de Minnesota, el Perfil de los Estados de Ánimo, las calificaciones de los compañeros, las autocalificaciones y los diagnósticos psiquiátricos).

Lista de Verificación de Adjetivos para la Depresión Estado-Rasgo (ST-DACL). Esos instrumentos breves (de dos a tres minutos) (de B. Lubin; Psychological Assessment Resources), los cuales constan de 32 a 34 adjetivos, fueron diseñados para medir sentimientos de disforia, tristeza y angustia psicológica. Cada una de las cinco formas (1, 2, A-B, C-D, Forma de Perfil) es administrada por la propia persona y proporciona una medida del estado de ánimo como estado (describe cómo se siente usted el día de la prueba) y del estado de ánimo como rasgo (describe cómo se siente en general). La ST-DACL puede utilizarse para evaluar el progreso en la consejería o en la psicoterapia, como instrumento de detección para identificar a personas con niveles significativos de depresión, y como medida de resultado repetido del éxito de programas de intervención.

Listas de verificación de problemas

Se han diseñado varias listas de verificación para identificar problemas conductuales en los niños, siendo una de las más antiguas la Lista Mooney de Verificación de Problemas. Uno de los instrumentos de este tipo citados con mayor frecuencia es la Lista de Verificación de la Conducta Infantil (CBCL). Al igual que la Lista Mooney de Verificación de Problemas, la CBLC es un *instrumento de banda amplia* que proporciona una perspectiva bastante incluyente del funcionamiento social, conductual y emocional. Otro ejemplo de una lista de verificación de problemas de banda amplia es la Lista de Verificación de Problemas de Conducta, revisada (RBPC). A diferencia de la lista Mooney, que es un *instrumento de autorreporte*, las dos últimas listas de verificación citadas son *instrumentos de informantes* que son llenados por un padre de familia o maestro. Hablando en términos estrictos, son escalas de calificación más que listas de verificación, ya que las respuestas se hacen sobre categorías múltiples.

Lista de Verificación de la Conducta Infantil. Este instrumento fue diseñado para evaluar los problemas y las competencias conductuales de los niños según el informe de los padres y de

otras personas que conozcan bien al niño. La versión de los padres de la CBCL consta de 118 reactivos de conducta problemática que se califican en una escala de cero (conducta que "no es cierta" del niño), uno (conducta "en ocasiones o algo cierta" del niño) y dos (conducta "muy cierta o a menudo cierta" del niño). Las calificaciones en los reactivos de competencia social se suman como subcalificaciones de Actividades, Social y Escuela.

La CBCL se estandarizó en 1981 en 1,300 estudiantes del área de Washington, D.C., y en el manual se proporcionan normas separadas para género y tres niveles de edad (4-5, 6-11, 12-16 años) en ocho a nueve factores (Achenbach y Edelbrock, 1983). Las normas arrojan seis diferentes perfiles de conducta infantil en ocho a nueve factores; se agrupan en síndromes de exteriorización, de interiorización y mixtos. Los coeficientes de confiabilidad test-retest en las variables de problemas de conducta y competencia social van de moderados a altos, mientras que los de los índices de acuerdo de los padres son mixtos. Se ha obtenido una cantidad sustancial de datos de validez para la CBCL. Por ejemplo, sus calificaciones tienen una correlación significativa con calificaciones en instrumentos similares como la Escala Conners de Calificación de los Padres (Conners, 1973; Conners y Barkley, 1985) y la Lista de Verificación de Problemas de Conducta, revisada (Quay y Peterson, 1983).

Formato de Informe del Maestro.[2] Una versión paralela de la Lista de Verificación de la Conducta Infantil, el Formato de Informe del Maestro (TRF) (Achenbach y Edelbrock, 1986), es llenada por los maestros o por sus ayudantes. La TRF proporciona una imagen de las conductas problemáticas y adaptativas de los niños en los escenarios escolares. Las personas que responden indican en una escala de tres puntos (no es cierto, algo o en ocasiones cierto, cierto muy a menudo) con cuánta frecuencia ocurrieron conductas específicas en los dos meses previos. El desempeño académico del niño se califica en una escala de cinco puntos (de "calificación muy por abajo" a "calificación muy por arriba"), y cuatro reactivos concernientes al funcionamiento conductual adaptativo se califican en una escala de siete puntos (de "mucho menos" a "mucho más"). La TRF se estandarizó en principio con una muestra de niños varones de seis a once años de edad, pero también se determinaron normas en otros grupos de niños. Los datos de confiabilidad y validez presentados para la TRF parecen ser satisfactorios (Edelbrock y Achenbach, 1984). Por ejemplo, las comparaciones entre calificaciones en la TRF de grupos de niños clínicos y no clínicos, además de una comparación de niños de grupos regulares con niños que recibían educación especial, han arrojado resultados significativos. También se ha encontrado que las correlaciones de las calificaciones de los niños en la TRF con sus conductas observadas son significativas (Edelbrock, 1988). Asimismo, se dispone de un Formato de Autorreporte Juvenil (YSR) de la CBCL diseñado para muchachos y muchachas de 11 a 18 años de edad (Achenbach y Edelbrock, 1987). Tanto la TRF como la YSR han recibido altas notas de los revisores como instrumentos para documentar las conductas problemáticas de niños y adolescentes (Christenson, 1992; Elliott y Busse, 1992). Sin embargo, los usuarios de esos instrumentos deben advertir que aunque pueden contribuir en los procesos clínicos de entrevista y toma de decisiones, no son instrumentos adecuados para utilizarse por sí solos con propósitos de diagnóstico o clasificación.

Lista de Verificación de Problemas de Conducta, revisada (RBPC). Similar a las listas Mooney de verificación, este instrumento de 89 reactivos (PAR) fue diseñado para identificar problemas de conducta en individuos de 5 a 18 años (Quay y Peterson, 1983). Se ha utilizado para detectar problemas de conducta en las escuelas, como auxiliar en el diagnóstico clínico, para

[2]La Lista de Verificación de la Conducta Infantil, el Formato de Informe del Maestro y el Formato de Autorreporte Juvenil pueden obtenerse de T. M. Achenbach y C. Edelbrock, Departamento de Psiquiatría, Universidad de Vermont.

medir el cambio conductual asociado con intervenciones psicológicas o farmacológicas, como parte de una batería para clasificar a infractores juveniles, y para seleccionar muestras de investigación sobre trastornos de conducta en niños y adolescentes. Puede ser llenada por un maestro, un padre de familia u otro observador aproximadamente en 20 minutos y se califica en seis subescalas: trastorno de conducta, agresión socializada, problemas de atención-inmadurez, ansiedad-alejamiento, conducta psicótica y tensión motriz-exceso. Se dispone de normas de calificación *T* basadas en las calificaciones de los maestros para los grados K a 12. Los coeficientes de confiabilidad entre calificadores para las seis subescalas van de moderados a altos, pero la confiabilidad test-retest es algo menor. El análisis de la validez de constructo de la RBPC indica que representa un consenso de lo que se sabe acerca de la conducta inadaptada del niño.

Listas de verificación de síntomas

Las listas de verificación de síntomas, como la Serie de Listas de Verificación del Estado Mental y la Serie Derogatis de Listas de Verificación de Síntomas, tienen una orientación más clínica que las listas de verificación de adjetivos o de problemas de conducta. Cada una de las dos listas de verificación del estado mental consta de 120 reactivos del tipo incluido en un examen integral del estado mental de un adulto: problemas presentados, datos de canalización, datos demográficos, estado mental, función de personalidad y síntomas, diagnóstico y disposición.

El instrumento clínico más popular en la Serie Derogatis de Listas de Verificación de Síntomas es la Lista de Verificación de Síntomas 90, revisada (SCL-90-R) (Derogatis, 1994; NCS Assessments). Los profesionales de la salud mental pueden aplicar la SCL-90-R en 12 a 15 minutos para evaluar a los pacientes psiquiátricos adolescentes o adultos en el momento de ingreso, detectar problemas psicológicos, supervisar el progreso o los cambios del paciente durante el tratamiento y evaluar los resultados posteriores al tratamiento. La SCL-90-R se califica en nueve dimensiones principales de síntomas: somatización, obsesivo compulsivo, sensibilidad interpersonal, depresión, ansiedad, hostilidad, ansiedad fóbica, ideación paranoide y psicoticismo. Tres medidas de estrés indican el nivel o la profundidad de un trastorno, la intensidad de los síntomas y el número de síntomas informados por el paciente. Se dispone de normas para adolescentes y adultos no pacientes y para pacientes psiquiátricos externos e internos.

ESTRATEGIAS PARA ELABORAR ESCALAS DE CALIFICACIÓN

Las escalas de calificación, las cuales fueron introducidas como instrumentos de investigación psicológica por Francis Galton durante la última parte del siglo XIX, son dispositivos de evaluación populares en los contextos clínico, escolar, laboral, deportivo y de entretenimiento. Las calificaciones pueden ser hechas por el calificado (la persona a la que se va a calificar) o por otro calificador. Las escalas de calificación se consideran, por lo general, menos precisas que los inventarios de personalidad y más superficiales que las técnicas proyectivas. Sea correcta o no esta percepción, las escalas de calificación tienen la doble ventaja de la economía y la versatilidad en su elaboración y calificación.

Una alternativa para una escala de calificación es una escala de rango, en la cual los individuos asignan rangos de 1 a *n* a *n* gente, objetos o acontecimientos (vea la sección de preguntas y ejercicios, punto 6). Aunque la elaboración de los instrumentos de rango es muy sencilla, a menudo su uso es engorroso: las personas que asignan los rangos pueden tener dificultades para efectuar el gran número de comparaciones $[n(n - 1)/2]$ requerido por el procedimiento de asignación de rangos. Calificar las respuestas a un instrumento de asignación de rangos no es particularmente difícil (vea el capítulo 3), pero el análisis estadístico de los resultados plantea algunos problemas.

No es muy difícil elaborar una escala de calificación. Todo lo que necesitamos hacer es designar o definir los objetos a ser evaluados, los atributos o características de los objetos a ser calificados, y las categorías (anclas) o el continuo sobre el que se harán las calificaciones. Sin embargo, para hacer un buen trabajo en la elaboración de la escala debemos tener primero una buena comprensión de las características particulares que van a calificarse y de los diversos formatos existentes para formular las calificaciones.

Entre las estrategias que se han seguido para elaborar escalas de calificación se encuentran la estrategia racional-teórica (deductiva), la estrategia de consistencia-interna (inductiva), y la estrategia de grupos-criterio (empírica). Se sigue una *estrategia-racional teórica* cuando la persona que elabora la escala se adhiere a los preceptos de una teoría particular o decide de otra manera qué parece razonable o lógico incluir en la escala. Por ejemplo, al elaborar una escala de calificación para medir ciertos aspectos de la personalidad, un seguidor de esta estrategia debe estar familiarizado con la investigación y la teoría de la personalidad y ser también un buen razonador lógico. En contraste con el proceso de razonamiento deductivo empleado en la estrategia racional-teórica, la *estrategia de consistencia-interna* hace uso de los resultados empíricos de la investigación empleando los métodos estadísticos de correlación y análisis factorial para seleccionar los reactivos que serán incluidos en la escala. Por último, la *estrategia de grupos-criterio* consiste en seleccionar o retener aquellos reactivos que diferencian entre dos o más grupos criterio de gente. Al planear una escala de calificación clínica, uno o más de los grupos criterio consta de pacientes a los que se les ha diagnosticado algún trastorno.

Esas tres estrategias para elaborar escalas de calificación no son, por supuesto, mutuamente excluyentes: dos de ellas o las tres pueden ser empleadas en alguna circunstancia en el proceso de elaborar una escala de calificación particular. Además, las estrategias no se limitan a las escalas de calificación. La elaboración de listas de verificación, de inventarios de personalidad, de técnicas proyectivas y de otros dispositivos para la evaluación de la personalidad puede basarse en una o más de esas estrategias.

TIPOS DE ESCALAS DE CALIFICACIÓN

Así como existen varias estrategias para elaborar escalas de calificación, existen diferentes formatos para presentar y responder a los reactivos en esos instrumentos. Entre los formatos se encuentran los siguientes tipos de escalas: numéricas, de analogía visual, de diferencial semántico, de calificación gráfica, de calificación estándar, con respaldo conductual y de elección forzada.

Escala numérica

En este tipo de escala de calificación se asigna a una persona, a un objeto o acontecimiento, uno de varios números correspondientes a las descripciones particulares de las características calificadas. Todo lo que se requiere es que las calificaciones se den en una escala ordenada en la cual se asignan diferentes valores numéricos a diferentes localizaciones. El formato 16.3 es ejemplo de un instrumento con 15 escalas numéricas de calificación que pueden usarse para calificarse uno mismo o calificar a alguien más. Las respuestas son calificadas en cinco variables de personalidad: agradabilidad, escrupulosidad, extroversión, neuroticismo y apertura a la experiencia. Las calificaciones en cada variable fluctúan entre 0 y 18.

Escalas unipolares y bipolares

Las escalas de calificación numérica, y muchos otros tipos de escalas, pueden ser unipolares o bipolares. En una *escala unipolar*, el atributo a calificar (por ejemplo, agresividad) es visto co-

FORMATO 16.3 Escala de calificación de personalidad de cinco variables

Instrucciones: Para cada reactivo, marque el número entre el par de adjetivos correspondientes a su descripción de usted mismo.

1.	afectuoso	1 2 3 4 5 6 7	reservado
2.	tranquilo	1 2 3 4 5 6 7	preocupado
3.	cuidadoso	1 2 3 4 5 6 7	descuidado
4.	conformista	1 2 3 4 5 6 7	independiente
5.	desorganizado	1 2 3 4 5 6 7	bien organizado
6.	realista	1 2 3 4 5 6 7	imaginativo
7.	ama la diversión	1 2 3 4 5 6 7	sobrio
8.	servicial	1 2 3 4 5 6 7	poco cooperativo
9.	inseguro	1 2 3 4 5 6 7	seguro
10.	prefiere la rutina	1 2 3 4 5 6 7	prefiere la variedad
11.	retraído	1 2 3 4 5 6 7	sociable
12.	despiadado	1 2 3 4 5 6 7	bondadoso
13.	autodisciplinado	1 2 3 4 5 6 7	con pobre voluntad
14.	autocompasivo	1 2 3 4 5 6 7	autosatisfecho
15.	suspicaz	1 2 3 4 5 6 7	confiado

mo unidimensional y, por ende, se considera que se incrementa de un mínimo a una cantidad máxima; los valores de la escala (anclas) son una serie de enteros crecientes. Por ejemplo, las anclas en una escala de cinco puntos pueden ser 0, 1, 2, 3, 4 o 1, 2, 3, 4, 5. En una *escala bipolar*, se considera que el atributo calificado varía en dos direcciones (por ejemplo, sumiso-agresivo); en consecuencia, la mitad de la escala se representa como 0 y los dos extremos (polos) son los enteros máximos negativo y positivo. Por ejemplo, las categorías numéricas en una escala bipolar de cinco puntos son $-2, -1, 0, 1, 2$.

Una escala unipolar se califica, por lo general, dando cero puntos a las calificaciones en la categoría correspondiente a la menor cantidad del atributo calificado, y $c - 1$ puntos, donde c es el número de categorías de calificación, a las calificaciones en la categoría que representa la cantidad más alta del atributo. Luego pueden sumarse las calificaciones a los reactivos para obtener una calificación parcial en un grupo particular de reactivos o una calificación total en el instrumento entero.

La calificación de las puntuaciones en las escalas bipolares implica dos pasos: primero, la misma cantidad de puntos (de 0 a $c - 1$) como en una escala unipolar se asigna para las calificaciones sucesivas, de la categoría más baja a la más alta; y luego se restan $(c - 1)/2$ puntos de cada uno de los puntos de categoría asignados en el paso 1. Por ejemplo, cuando existen cinco categorías bipolares, primero se asignan 0, 1, 2, 3 y 4 puntos a las categorías sucesivas de calificación. Al restar $(5 - 1)/2 = 2$ de cada uno de esos valores obtenemos -2, -1. 0, 1 y 2, los cuales son las puntuaciones del reactivo para calificaciones asignadas en las cinco categorías sucesivas de la escala bipolar. Al igual que con las puntuaciones en la escala unipolar, las calificaciones resultantes del reactivo pueden sumarse luego para proporcionar una calificación parcial o total.

Diferencial semántico

Un tipo de escala numérica de calificación que se ha empleado con frecuencia en la investigación sobre psicología social y de la personalidad es el *diferencial semántico*. Osgood, Suci y Tannenbaum (1957) idearon este método para sus estudios de los significados connotativos (personales) que conceptos como PADRE, MADRE, ENFERMEDAD, PECADO, ODIO y AMOR tienen para diferentes personas. Cuando se presenta un instrumento de diferencial semántico, la persona califica una serie de conceptos en varias escalas de adjetivos bipolares de siete puntos. Por ejemplo, el concepto MADRE puede ser calificado colocando una marca en el segmento apropiado de la línea en cada una de las tres siguientes escalas:

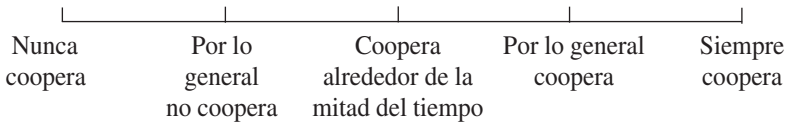
MALO ____:____:____:____:____:____:____ BUENO
DÉBIL ____:____:____:____:____:____:____ FUERTE
LENTO ____:____:____:____:____:____:____ RÁPIDO

Una vez que todos los conceptos de interés han sido calificados en las diversas escalas, las respuestas a cada concepto se califican en varias *dimensiones semánticas* y se comparan con las respuestas a los conceptos restantes. Las principales dimensiones (semánticas) del significado connotativo que han sido determinadas por el análisis factorial de las calificaciones de una serie de conceptos en un gran número de esas escalas de adjetivos son *evaluación, potencia* y *actividad*. Luego puede elaborarse un *espacio semántico* al graficar las calificaciones de una persona sobre los conceptos calificados en cada una de esas tres dimensiones. Se supone que los conceptos que se mantienen cercanos entre sí en el espacio semántico tienen significados connotativos similares para el calificador.

Escala de calificación gráfica

Otro popular tipo de escala de calificación es la *escala de calificación gráfica*, un ejemplo de la cual es:

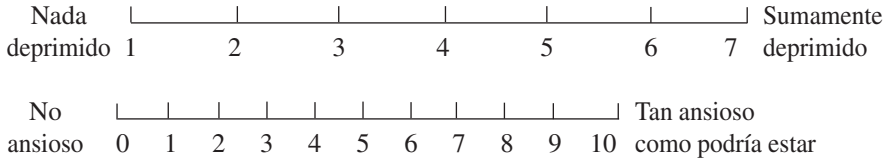
¿Qué tan bien coopera esta persona en un grupo?



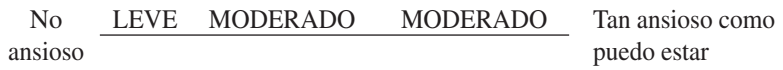
El calificador escribe una X o coloca otra marca en cada una de una serie de líneas, como la del ejemplo precedente, que contienen palabras o frases descriptivas correspondientes a cierta característica o rasgo. Por lo regular, una descripción del menor grado de la característica se presenta en el extremo izquierdo de la línea, una descripción del grado más alto de la característica se presenta al extremo derecho, y las descripciones referentes a los grados intermedios de la característica se presentan en los puntos intermedios de la línea.

Escala de analogía visual

En los contextos clínicos a menudo es difícil determinar la intensidad de la experiencia subjetiva de un paciente (de dolor, ansiedad, anhelo de una sustancia y situaciones similares). La *escala de analogía visual* es una técnica empleada para estimar la intensidad de dichas experiencias (Wewers y Low, 1990). Por ejemplo, puede indicarse al paciente que señale o marque el lugar en la línea que corresponde a la intensidad de la ansiedad o dolor que experimenta en ese momento. Puede pedirse a un niño pequeño que señale la ilustración de un rostro, en una serie graduada de rostros sonrientes y ceñudos, que mejor indique cómo se siente. Las siguientes escalas son ejemplos de escalas de analogía visual con anclas numéricas.



Ejemplo de una escala de analogía visual con anclas verbales es



Es posible aplicar de manera periódica escalas de analogía visual como ésta para medir cambios en los sentimientos o estados de ánimo (por ejemplo, las Escalas Stern de Analogía Visual del Estado de Ánimo) a lo largo del tiempo, pero tienen limitaciones.

Esta técnica es quizá más precisa que pedir simplemente a los pacientes que digan en sus propias palabras cómo se sienten, pero algunos pacientes tienen dificultad para entender el procedimiento y representar experiencias subjetivas como el dolor, la ansiedad y la depresión en escalas de analogía visual.

Escala de calificación estándar

En una *escala de calificación estándar*, el calificador proporciona o se le proporciona un conjunto de estándares para evaluar a las personas que se califican (los *calificados*). Ejemplo de una escala de calificación estándar es la *escala persona a persona*, la cual se elabora para calificar a individuos en un rasgo especificado, como la capacidad de liderazgo. Se pide al calificador pensar en cinco personas que se localizan en diferentes puntos a lo largo de un continuo hipotético de capacidad de liderazgo. Luego el calificador compara a cada calificado con esos cinco individuos e indica cuál de ellos se parece más al calificado en la capacidad de liderazgo.

Escalas con respaldo conductual

Desarrolladas por Smith y Kendall (1963) y basadas en la *técnica de incidentes críticos* de Flanagan (1954), las escalas con respaldo conductual representan un intento por lograr que la terminología de las escalas de calificación sea más descriptiva de la conducta real y, por ende, más objetiva. Como es comprensible, términos como *ansiedad, autoconfianza, agresividad* y otros sustantivos y adjetivos usados en las escalas de calificación tradicionales orientadas a los rasgos pueden ser interpretados de manera diferente por calificadores distintos. Esto es cierto en particular cuando los calificadores reciben poca o ninguna capacitación sobre cómo interpretar los términos. En el formato 16.4 se presenta una ilustración irónica de una escala con respaldo conductual para calificar los factores de desempeño de cinco empleados.

La elaboración de una escala de calificación con respaldo conductual comienza por convocar a un grupo de individuos que posean conocimiento experto sobre un trabajo o situación en particular. Luego, por medio del análisis y la deliberación, esos individuos intentan alcanzar un consenso sobre una serie de incidentes críticos conductualmente descriptivos, a partir de los cuales pueda elaborarse una escala de calificación objetiva y muy confiable. Las descripciones conductuales que sobreviven a repetidas evaluaciones por parte del grupo o de otros grupos pueden prepararse entonces como una serie de reactivos a ser calificados. Podríamos esperar que el énfasis en la conducta observable y el esfuerzo concentrado del grupo por idear escalas con respaldo conductual consiguieran que éstas fueran superiores en lo psicométrico a otros tipos de escalas de calificación. Además, el hecho de que el proceso de elaboración de la escala requiera partici-

FORMATO 16.4 Guía para la valoración del desempeño de los empleados

	GRADOS DE DESEMPEÑO				
FACTORES DE DESEMPEÑO	*Excede con mucho los requerimientos del empleo*	*Excede los requerimientos del empleo*	*Cumple los requerimientos del empleo*	*Necesita mejorar*	*No cumple los requerimientos mínimos*
Calidad	Salta edificios altos con un solo impulso	Debe correr para impulsarse para saltar edificios altos	Sólo puede saltar un edificio bajo o mediano sin punta	Se estrella contra los edificios cuando intenta saltarlos	No puede reconocer los edificios, mucho menos saltarlos
Puntualidad	Es más rápido que una bala	Es tan rápido como una bala	No es tan rápido como una bala	¿Creería que es como una bala lenta?	Se lastima con las balas cuando intenta disparar un arma
Iniciativa	Es más fuerte que una locomotora	Es más fuerte que un elefante macho	Es más fuerte que un toro	Le dispara al toro	Huele como un toro
Adaptabilidad	Camina sobre el agua de manera consistente	Camina sobre el agua en emergencias	Lava con agua	Bebe agua	Pasa el agua en las emergencias
Comunicación	Habla con Dios	Habla con los ángeles	Habla consigo mismo	Discute consigo mismo	Pierde esas discusiones

Adaptado de *The Industrial Organization Psychologist,* 1980, *17*(4), p. 22, y utilizado con autorización.

pación y consenso de grupo, y por ende mayor probabilidad de aceptación por el grupo, podría parecer una ventaja. Sin embargo, los resultados de la investigación indican que las escalas de calificación con respaldo conductual no por fuerza son superiores a las escalas de calificación gráfica (Kinicki y Bannister, 1988).

Dos variaciones de las escalas con respaldo conductual son las *escalas de expectativa conductual* (*BES*) y las *escalas de observación conductual* (*BOS*). En las escalas BES, las conductas cruciales se califican en términos de expectativas más que como conductas reales. Las calificaciones en las escalas BOS se asignan en términos de la frecuencia (nunca, rara vez, en ocasiones, por lo general, siempre) con que cada una de un conjunto de conductas críticas es observada durante un periodo especificado. Algunos investigadores han concluido que en contextos de empleo el método BOS es preferible al BARS (Wiersma y Latham, 1986).

Escala de elección forzada

En una *escala de calificación de elección forzada* se presentan al calificador dos o más descripciones y se le pide indicar cuál caracteriza mejor a la persona calificada. Si hay tres o más descripciones, también puede pedirse a los calificadores indicar cuál es la menos descriptiva del calificado. Por ejemplo, en un reactivo que contiene cuatro descripciones, dos son igualmente deseables y dos igualmente indeseables. Se pide al calificador seleccionar la afirmación más descriptiva y la menos descriptiva del calificado. Sólo una afirmación deseable y una indeseable discriminan en realidad entre los calificados altos y bajos en el criterio, pero se supone que los calificadores no saben cuáles son esas afirmaciones. Ejemplo hipotético de un reactivo de elección forzada de cuatro afirmaciones para calificar el liderazgo es:

____ Asume la responsabilidad con facilidad.
____ No sabe cómo o cuándo delegar.
____ Tiene muchas sugerencias constructivas que ofrecer.
____ No escucha las sugerencias de otros.

(Lector: ¿puede señalar cuál afirmación se identifica como "deseable" y cuál como "indeseable"?)

Los calificadores en ocasiones encuentran que el formato de elección forzada es engorroso, pero se considera más justo que la técnica de calificación persona a persona. La técnica de elección forzada también tiene la ventaja de controlar ciertos errores en la calificación, como los errores constantes, el efecto de halo, el error de contraste y el de proximidad.

PROBLEMAS CON LAS CALIFICACIONES

Al asignar calificaciones puede cometerse una variedad de errores, entre los cuales se encuentran los errores constantes, el efecto de halo, el error de contraste y el de proximidad. No todos los calificadores son igualmente proclives a cometer esos errores; como con cualquier otro método de evaluación, eso depende de la capacidad de observación y de la experiencia y personalidad del calificador, y presumiblemente de influencias hereditarias en sus habilidades interpersonales, capacidad de percepción y libertad de sesgos de juicio.

Errores en la calificación

Los *errores constantes* ocurren cuando las calificaciones asignadas en la categoría promedio son más altas (*error de generosidad* o *indulgencia*), más bajas (*error de severidad*) o más frecuentes

(*error de tendencia central*) de lo que deben ser. Todos los calificadores son susceptibles al *error de ambigüedad* de no lograr interpretar los reactivos de manera correcta porque están mal planteados, porque se proporciona información insuficiente o porque las anclas de la escala no se describen o se colocan de manera apropiada.

Otro tipo de error de calificación, el *efecto de halo*, ocurre cuando los calificadores muestran la tendencia a responder con base en una impresión general del calificado o a generalizar en exceso dando calificaciones favorables a todos los rasgos sólo porque el calificado sobresale en uno o dos. El efecto de halo también puede ser negativo, en cuyo caso una mala característica afecta las calificaciones en todas las otras características. En relación con el efecto de halo, existe el *error lógico* de asignar calificaciones similares en características que el calificador percibe como lógicamente relacionadas.

El término *error de contraste* ha sido empleado al menos en dos sentidos. En un sentido se refiere a la tendencia a asignar calificaciones más altas de lo justificado si el calificado inmediatamente precedente recibió una calificación muy baja, o a asignar una calificación más baja de lo justificado si el calificado anterior recibió una calificación muy alta. En un segundo sentido, el error de contraste se refiere a la tendencia de un calificador a comparar o contrastar al calificado consigo mismo al asignar calificaciones en ciertas conductas o rasgos.

Un *error de proximidad* ocurre cuando el calificador tiende a asignar calificaciones similares a una persona en reactivos que están colocados juntos en la página impresa. De igual modo, si una persona recibe una calificación consistentemente alta, baja o promedio en la mayoría de un conjunto de reactivos que están cercanos en la página impresa, puede recibir calificaciones similares en otros reactivos localizados cerca de ellos. Otro factor de proximidad, el *error de desempeño más reciente* ocurre cuando un calificado es juzgado sobre la base de su conducta más reciente, en lugar de sobre una muestra más representativa de la conducta.

Los errores en la calificación también pueden ocurrir cuando los calificadores poseen información inadecuada acerca de las personas calificadas. En consecuencia, los calificadores pueden recibir una clara influencia de comunicaciones irrelevantes o incorrectas acerca del calificado y atribuir mucha importancia a detalles insignificantes concernientes al mismo. Al enfrentarse a un conocimiento insuficiente acerca del calificado, los calificadores pueden retroceder hacia estereotipos acerca de la naturaleza humana, recordar sólo la información que confirma sus creencias acerca del calificado y de la gente en general, y a ser más dirigidos por sus sentimientos que por la información correcta. Los calificadores también pueden cometer el *error fundamental de atribución* de interpretar que la conducta del calificado en la situación de calificación se debe a factores internos o disposicionales más que a la propia naturaleza de la situación de calificación.

Mejoramiento de las calificaciones

No es fácil formular juicios confiables y válidos acerca de la gente en la mejor de las circunstancias, y sobre todo cuando las conductas o características están mal definidas o son muy subjetivas. No sólo es probable que los sesgos personales afecten las calificaciones, sino que a menudo los calificadores no tienen la familiaridad suficiente con los calificados como para emitir juicios precisos. La capacitación sobre cómo asignar las calificaciones más objetivas —estar al tanto de los varios tipos de errores que pueden ocurrir en la calificación, familiarizarse con las personas y los rasgos que van a ser calificados, y omitir los reactivos que el calificador sienta no está preparado para juzgar— puede mejorar la precisión de las calificaciones (Stamoulis y Hauenstein, 1993; Sulsky y Day, 1994). La combinación de las respuestas de varios calificadores también

puede equilibrar los sesgos de respuesta de los calificadores individuales. Es posible obtener mayor confiabilidad y validez de las calificaciones al diseñar los reactivos con más cuidado y plantearlos en la terminología conductual precisa; al disponerlos en las hojas de calificación de forma que puedan ser leídos y calificados con mayor facilidad y precisión, y al asegurarse de que los reactivos individuales y el formato de calificación como un todo no sean excesivamente largos.

La investigación sobre calificaciones del trabajo ha demostrado que las calificaciones más confiables son las dadas por los pares del calificado (Imada, 1982; Wexley y Klimoski, 1984). Las calificaciones de subordinados, superiores, pares y la persona no siempre concuerdan, pero la combinación de las calificaciones de esas cuatro fuentes puede resultar en una mayor confiabilidad y validez que de cualquiera de las fuentes por sí mismas (Harris y Schaubroeck, 1988). Por último, una atención cuidadosa al diseño de las escalas de calificación, definiendo los puntos (*anclas*) con claridad mediante la descripción conductual precisa de las características a ser calificadas, contribuye a garantizar la validez de las calificaciones.

ESCALAS DE CALIFICACIÓN ESTANDARIZADAS

La gran mayoría de las escalas de calificación no son estandarizadas, son instrumentos elaborados con propósitos especiales, diseñados para investigaciones particulares. No obstante, en el mercado pueden encontrarse muchas escalas estandarizadas para calificar la conducta y los rasgos de personalidad de niños y adultos. Las escalas para calificar el estado de desarrollo y las conductas de los niños con retraso mental, discapacidad de aprendizaje, perturbación emocional e impedimentos físicos son muy populares. También son de gran uso las escalas para calificar la ansiedad, la depresión, la hostilidad y otros síntomas clínicos.

Los investigadores en el campo del desarrollo infantil, la educación especial (por ejemplo, autismo, TDAH, deterioros del habla y el lenguaje, retraso mental), y la psicología escolar en particular, han elaborado docenas de instrumentos de calificación para evaluar los cambios conductuales que resultan de intervenciones educativas, terapéuticas y de otros programas específicos. Muchos de esos instrumentos están orientados hacia la evaluación conductual, mientras que otros tienen una orientación rasgo-factor y algunos más fueron desarrollados en un contexto psicodinámico, psiquiátrico y, por ende, contienen terminología asociada. Además, muchas entrevistas estandarizadas e instrumentos de observación involucran la calificación de la conducta y la personalidad, de ahí que se constituyan en parte de escalas de calificación.

Se dispone de informes basados en la computadora para muchas escalas de calificación y listas de verificación distribuidas de manera comercial. Además, una serie de escalas de calificación y listas de verificación pueden ser aplicadas por computadora. Por ejemplo, existen versiones aplicadas por computadora de escalas de calificación administradas por clínicos, algunas de las cuales emplean *respuesta por voz interactiva* (IVR), para la evaluación de la ansiedad, la depresión, el trastorno obsesivo-compulsivo y la fobia social. En su revisión de las escalas de calificación clínica aplicadas por computadora, Kobak, Greist, Jefferson y Katzelnick (1996) concluyeron que los pacientes eran más honestos, por lo general, con la computadora que con otros métodos de presentación, y que a menudo la preferían cuando revelaban información delicada acerca de suicidio, abuso de alcohol o drogas, conducta sexual y síntomas relacionados con el VIH. Los revisores concluyeron que, cuando se usan con directrices éticas establecidas, las computadoras son confiables, económicas, accesibles y permiten un uso eficiente del tiempo en la evaluación de los síntomas psiquiátricos.

CLASIFICACIONES Q Y LA PRUEBA REP

Las clasificaciones Q son similares a las escalas de calificación, pero también poseen ciertos rasgos de las listas de verificación. La *técnica de clasificación Q*, iniciada por Stephenson (1953), requiere que el individuo clasifique un conjunto de afirmaciones descriptivas en una serie de pilas que van de lo "más característico" a lo "menos característico" de sí mismo o de un conocido. Se pide a la persona que ordene las afirmaciones de modo que un número especificado de éstas quede en cada pila y produzca una distribución normal de afirmaciones entre las pilas.

Las afirmaciones de las clasificaciones Q pueden prepararse de manera específica para cierta investigación, pero se dispone de grupos de afirmaciones estándar. Un conjunto distribuido de manera comercial, la Clasificación Q de California, revisada (Conjunto para los Adultos), consta de 100 tarjetas que contienen afirmaciones descriptivas de personalidad; también puede encontrarse en Consulting Psychologist Press un Conjunto para Niños.

Ciertas investigaciones sobre los cambios en el autoconcepto resultantes de la psicoterapia o de otras intervenciones han requerido que los sujetos de la investigación realicen clasificaciones Q previas y posteriores de una serie de afirmaciones que describen sus sentimientos y actitudes (por ejemplo, Rogers y Dymond, 1954). Cuando las clasificaciones del yo verdadero y del yo ideal son más parecidas después de la intervención de lo que eran antes de ésta, puede concluirse que la experiencia de intervención fue efectiva.

En lugar de pedir a las personas que clasifiquen las afirmaciones, puede pedírseles que clasifiquen a un conjunto de individuos en varias categorías. Un ejemplo de este enfoque es la Prueba de Repertorio de Construcción de Papeles (Rep). De acuerdo con Kelly (1955), las personas se parecen a los científicos en que conceptualizan o categorizan sus experiencias de una manera que les parece lógica. Por desgracia, mucha gente percibe o construye el mundo de manera incorrecta y, por ende, desarrolla un sistema erróneo de constructos. El objetivo de la prueba Rep es identificar el sistema de constructos personales que una persona utiliza para interpretar sus experiencias. Al presentar la prueba Rep, el examinado clasifica a las personas que son importantes para él de ciertas maneras en varias categorías conceptuales que selecciona por sí mismo. El desempeño en la prueba Rep es analizado al advertir cuántos constructos son usados por el individuo, cuáles son éstos, qué características de la gente son enfatizadas por esos constructos (físicos, sociales, etc.), y qué personas son más parecidas o más diferentes al sujeto. La interpretación de los resultados de la prueba Rep en términos del *sistema de constructos personales* del individuo, el cual sirve como marco interno de referencia para percibir y entender el mundo, es un proceso laborioso y subjetivo. Este hecho, aunado a la escasa evidencia en favor de la validez de la prueba Rep, ha dado por resultado un uso infrecuente de ésta en los programas clínicos y de investigación.

RESUMEN

Las listas de verificación y las escalas de calificación se utilizan en contextos educativos, ocupacionales y clínicos para determinar si la gente posee ciertas características, rasgos o conductas deseables o indeseables. Para llenar una lista de verificación es necesario enfrentar una serie de decisiones dicotómicas (sí/no, cierto/falso, etc.), mientras que responder a escalas de calificación requiere una decisión evaluativa de categorías múltiples. Entre los muchos propósitos cumplidos por las listas de verificación y las escalas de calificación se encuentra proporcionar un registro objetivo de los resultados de observaciones y entrevistas. Esos instrumentos también pueden emplearse para determinar si ocurren cambios en una conducta como resultado de un tratamiento en particular, un programa educativo u otro procedimiento de intervención.

Las listas de verificación generalmente son muy sencillas de elaborar, pero resultan más objetivas cuando los reactivos tratan con conductas específicas. Aunque muchas listas de verificación son instrumentos caseros diseñados para una investigación específica o un propósito práctico, en el mercado se dispone de docenas de ellas. Las listas de verificación de problemas, conducta adaptativa, desarrollo, síntomas psiquiátricos y muchos otros rasgos se han usado de manera amplia con propósitos de diagnóstico e investigación en contextos educativos, clínicos e industrial-organizacionales.

Tres estrategias que se emplean al elaborar escalas de calificación son la racional-teórica, la de consistencia-interna y la de grupos-criterio. Las escalas de calificación válidas requieren que calificadores objetivos y sin sesgos emitan juicios (calificaciones) acerca de conductas, rasgos de personalidad y otras características de los individuos (calificados). Se ha utilizado una variedad de formatos al elaborar escalas de calificación, incluyendo los de tipo numérico, estándar (persona a persona), gráfico, diferencial semántico, con respaldo conductual, de elección forzada y de analogía visual. Cada tipo de escala tiene ventajas y desventajas, y cada escala es más útil para algunos propósitos que para otros.

Entre los muchos errores que se cometen al elaborar escalas de calificación están el error de ambigüedad, errores constantes (de indulgencia, severidad y tendencia central), el error de contraste, el error lógico, el error de proximidad y el efecto de halo. El procedimiento de calificación de elección forzada, en el cual se requiere que el calificador elija entre dos descripciones igualmente deseables y quizá también entre dos descripciones igualmente indeseables, controla algunos de esos errores, pero su uso es engorroso y a muchos calificadores les disgusta. Las calificaciones pueden ser transformadas a calificaciones estándar como un control estadístico para prevenir los errores constantes, pero quizá el procedimiento más efectivo para reducir los efectos de cualquier tipo de error en la calificación sea capacitar con cuidado a los calificadores y familiarizarlos con los diversos errores que pueden cometerse.

Cuando las escalas de calificación se elaboran con cuidado, se hacen tan objetivas como sea posible, y se capacita a los calificadores de manera concienzuda, pueden obtenerse coeficientes de confiabilidad del orden de .80 o incluso de .90. Promediar las calificaciones de varios calificadores también mejora el coeficiente de confiabilidad de una escala de calificación.

Las clasificaciones Q son escalas de calificación modificadas en las cuales los individuos clasifican un conjunto de 100 tarjetas, o algo así, que contienen descripciones de personalidad en nueve pilas para formar una distribución normal de las afirmaciones entre las pilas. El procedimiento de clasificación Q ha sido empleado en estudios concernientes a la efectividad de la consejería psicológica y en otros contextos de investigación y aplicados. La Prueba de Repertorio de Construcción de Papeles (prueba Rep) fue diseñada por George Kelly para identificar el sistema de constructos personales de una persona a fin de determinar qué aspectos de la gente son enfatizados en el constructo y qué personas son más similares o diferentes de quien responde. La prueba Rep no ha sido usada de manera amplia con propósitos clínicos o de investigación, y en gran medida se desconoce su validez.

PREGUNTAS Y ACTIVIDADES

1. Consulte un diccionario o un compendio especializado y seleccione una muestra de 50 adjetivos referentes a rasgos o características personales. Forme una mezcla de términos positivos y menos positivos que no sean sinónimos o antónimos. Haga múltiples copias de la lista alfabetizada de los términos. Coloque una línea corta delante de cada adjetivo y presente la lista a una muestra de per-

sonas. Pídales que marquen cada adjetivo que crean las describe de manera general. Resuma los resultados comparándolos con lo que ya sabe acerca de las personas a partir de otros informes y observaciones.

2. Elabore una lista de verificación de 10 reactivos de conductas que sean sintomáticas de la depresión, y una segunda lista de verificación de conductas que sean sintomáticas de la ansiedad. Haga copias de estas dos listas y aplíquelas a doce personas. Califique las listas de verificación contando el número de reactivos marcados por quienes respondieron. Calcule e interprete la correlación entre las calificaciones de las personas en las dos listas de verificación.

3. Un problema con la literatura sobre la investigación de la conducta tipo A es que diferentes métodos de evaluación (por ejemplo, entrevista y cuestionario) no arrojan los mismos resultados. Aunque cuestionarios como la Encuesta de Actividad Jenkins son más eficientes que las entrevistas, Rosenman (1986) y otros han rechazado dichas medidas de autorreporte porque se supone que las personalidades tipo A tienen poco *insight* sobre su propia conducta. Una forma de probar esta hipótesis es comparar las calificaciones que la persona asigna a su propia conducta con calificaciones de ese comportamiento formuladas por observadores no sesgados. Con esto en mente, seleccione a unos cuantos individuos que parezcan ajustarse a la siguiente descripción de la personalidad tipo A:

> Un patrón de personalidad caracterizado por una combinación de conductas, incluyendo agresividad, competitividad, hostilidad, acciones rápidas y esfuerzo constante.

Aplique la lista de verificación del formato 16.1 a cada persona, y luego solicite a alguien que la conozca bien que llene la misma lista de verificación para describir a esa persona. Use un procedimiento estadístico apropiado para comparar las autocalificaciones con las calificaciones de los otros.

4. Califíquese en cada una de las siguientes características en una escala de 1 (“Considerablemente muy por abajo del promedio”) a 10 (“Considerablemente muy por arriba del promedio”).

____ **1.** habilidad para llevarse bien con los demás
____ **2.** habilidad atlética
____ **3.** cooperatividad
____ **4.** creatividad
____ **5.** nivel de energía
____ **6.** espíritu de servicio
____ **7.** inteligencia
____ **8.** habilidad de liderazgo
____ **9.** paciencia
____ **10.** sensatez
____ **11.** responsabilidad
____ **12.** sinceridad
____ **13.** previsión
____ **14.** tolerancia
____ **15.** integridad

Use el siguiente procedimiento para evaluar sus respuestas: sume sus calificaciones en las 15 características y divida la suma entre 15 para obtener la calificación promedio. Una calificación media “promedio” es 5.5, pero si usted es como la mayoría de los estudiantes su promedio será mayor. Este fenómeno de “mejor que el promedio”, el cual se relaciona con el grupo de respuesta de “deseabilidad social”, es una tendencia en que la mayoría de la gente se ve como mejor que el promedio.

5. En una escala de 1 a 10, donde 1 es la menor calificación y 10 la calificación más alta, califique cada uno de los siguientes adjetivos de acuerdo con qué tan descriptivos son de (a) su *yo verdadero* (la forma en que usted es en realidad), (b) su *yo ideal* (la forma en que le gustaría ser), y (c) otra *gente en general*.

	SU YO VERDADERO	SU YO IDEAL	OTRA GENTE EN GENERAL
valiente	____	____	____
cuidadoso	____	____	____
alegre	____	____	____
escrupuloso	____	____	____
considerado	____	____	____
cortés	____	____	____
creativo	____	____	____
confiable	____	____	____
vigoroso	____	____	____
amistoso	____	____	____
bien parecido	____	____	____
servicial	____	____	____
honesto	____	____	____
gracioso	____	____	____
inteligente	____	____	____
organizado	____	____	____
paciente	____	____	____
fuerte	____	____	____
estudioso	____	____	____
confiado	____	____	____

Evalúe sus respuestas mediante el siguiente procedimiento: calcule la suma de los valores absolutos de las diferencias entre las calificaciones asignadas a (a) su yo verdadero y su yo ideal, (b) su yo verdadero y los yo de la otra gente en general y (c) su yo ideal y los yo de la otra gente en general. Calcule el porcentaje del coeficiente de congruencia para cada una de las tres comparaciones dividiendo la suma entre 180 y restando el cociente resultante de 1. Entre más cercano sea el coeficiente de congruencia a 1.00, más similares son los dos yo. Interprete sus resultados en términos de la teoría del yo de Rogers o de la teoría del aprendizaje social.

6. Ordene, del 1 al 12, cada uno de los siguientes conjuntos de tres adjetivos en términos de qué tan descriptivo de su personalidad es cada conjunto. Un rango de 1 significa que los tres adjetivos lo describen de manera plena, y un rango de 12 que no lo describen.

____ **1.** iniciador, entusiasta y valeroso
____ **2.** estable, obstinado y bien organizado
____ **3.** intelectual, adaptable y listo
____ **4.** sensible, nutriente y compasivo
____ **5.** extrovertido, generoso y autoritario
____ **6.** crítico, exigente e inteligente
____ **7.** concertador, justo y sociable
____ **8.** reservado, fuerte y apasionado
____ **9.** honesto, impulsivo y optimista
____ **10.** ambicioso, trabajador y cauteloso
____ **11.** original, receptivo e independiente
____ **12.** gentil, sensible y creativo

De acuerdo con la astrología, las características de personalidad de un individuo son determinadas por el signo zodiacal de su fecha de nacimiento. Los 12 signos del zodiaco y las fechas correspondientes son las siguientes:

1. Aries: 21 de marzo a 19 de abril
2. Tauro: 20 de abril a 20 de mayo
3. Géminis: 21 de mayo a 21 de junio
4. Cáncer: 22 de junio a 22 de julio
5. Leo: 23 de julio a 22 de agosto
6. Virgo: 23 de agosto a 22 de septiembre
7. Libra: 23 de septiembre a 22 de octubre
8. Escorpión: 23 de octubre a 21 de noviembre
9. Sagitario: 22 de noviembre a 21 de diciembre
10. Capricornio: 22 de diciembre a 19 de enero
11. Acuario: 20 de enero a 18 de febrero
12. Piscis: 19 de febrero a 20 de marzo

¿Corresponde el número de su signo zodiacal con el número de la tríada de reactivos a la que le dio el rango de 1? Compare sus resultados con los de sus compañeros de clase, amigos y familiares. ¿Es ésta una prueba justa en relación con la validez del proceso de analizar la personalidad en términos de los signos del zodiaco? ¿Por qué sí o por qué no? ¿Cree usted en la astrología? Defienda su respuesta. [Adaptado de Balch, W. R. (1980). Testing the validity of astrology in class. *Teaching of Psychology*, 7(4), pp. 247-250.]

CAPÍTULO DIECISIETE

INVENTARIOS DE PERSONALIDAD

Las escalas de calificación y las listas de verificación han contribuido a la evaluación y comprensión de la personalidad humana, pero la mayoría de esos instrumentos se originaron en circunstancias diferentes a las que propiciaron la creación de los inventarios de personalidad analizados en este capítulo. Aunque una serie de escalas de calificación y listas de verificación han sido diseñadas para el psicodiagnóstico y otros propósitos clínicos, la mayoría de ellas fueron elaboradas para utilizarse en contextos educativos y de empleo. Algunos inventarios de personalidad también han sido diseñados y aplicados en esos contextos, pero los más populares se han concentrado en la identificación de trastornos emocionales y en el diagnóstico de la psicopatología en situaciones clínicas.

Este capítulo proporciona un panorama de los inventarios de personalidad de una sola calificación y de calificación múltiple diseñados para una variedad de propósitos aplicados y de investigación. Desde los primeros años del siglo XX se han elaborado cientos de inventarios de personalidad. Algunos fueron diseñados para aplicarse sobre todo en contextos educativos, otros en contextos clínicos y otros más en contextos ocupacionales. En este capítulo se presentan inventarios de diferentes tipos que se basan en diversas concepciones de la personalidad, pero más que ser una muestra representativa de todos los inventarios disponibles, sólo se exponen aquellos que están bien diseñados y que se han investigado de manera exhaustiva.

VERACIDAD, CONFIABILIDAD Y VALIDEZ

Los inventarios de personalidad constan de reactivos que atañen a las características personales, los pensamientos, sentimientos y la conducta. Al igual que en un inventario de intereses, una escala de calificación o una lista de verificación, en un inventario de personalidad los individuos marcan los reactivos que juzgan descriptivos de sí mismos o, en ciertos casos, de alguien a quien conocen bien.

Veracidad al responder

Al igual que con cualquier medida de lápiz y papel de las características y comportamientos humanos, un problema relacionado con los inventarios de personalidad tiene que ver con su validez. Como se ha enfatizado a lo largo de este texto, un instrumento psicométrico no puede proporcionar resultados válidos a menos que sea respondido de manera consistente y honesta. Y como muchos de los reactivos de los inventarios requieren que quienes responden admitan cosas que podrían desear no admitir, sino más bien presentarse de la manera más favorable, la cuestión de la veracidad en las respuestas es seria.

La veracidad al responder puede ser un problema grave en los inventarios de personalidad. Es probable que los individuos no estén dispuestos a decir la verdad o que ni siquiera sepan la verdad acerca de sí mismos y, en consecuencia, proporcionen información incorrecta. Es penosamente asombroso encontrar que la gente puede responder a los inventarios de personalidad de manera distorsionada cuando se le indica hacerlo. Pero sea por temor a ser descubiertos o por cualquier otra razón, las mentiras e imposturas en los inventarios psicológicos no son tan comunes en situaciones de orientación o ubicación laboral como podríamos sospechar (Schwab y Packard, 1973). Se han diseñado claves especiales para la validación de la calificación con el propósito de detectar simulaciones o imposturas en algunos inventarios. Las calificaciones obtenidas al aplicar esas claves no siempre revelan si las personas han sido descuidadas o mentirosas, pero permiten verificar la validez de los hallazgos.

El engaño intencional, ya sea aparentar ser peor (*mentir que es malo*) o mejor (*mentir que es bueno*) de lo que se es, no constituye el único factor que afecta la precisión de las respuestas a un inventario de personalidad. Las tendencias o grupos de respuesta, tales como conformidad, deseabilidad social, cautela excesiva y rigurosidad, también influyen en la validez de la calificación. De particular interés son los grupos de respuesta de *conformidad* (la tendencia a estar de acuerdo más que en desacuerdo cuando se duda) y de *deseabilidad social* (la tendencia a responder de una manera que sea más aceptable para la sociedad). Como con las mentiras de que se es bueno o se es malo, en algunos inventarios se han elaborado claves especiales de calificación para detectar o compensar esos grupos de respuesta. Por lo general, las calificaciones de una persona en esas *escalas de validez* se inspeccionan antes de evaluar las calificaciones en otras escalas (de contenido o de diagnóstico). Debido a que las calificaciones en las escalas de validez no por necesidad revelan la impostura y los grupos de respuesta, es mejor usar los inventarios de personalidad como auxiliares en la toma de decisiones sólo cuando los individuos no tengan nada que perder al responder de manera cuidadosa y con veracidad y no tengan nada que ganar al no hacerlo de esa manera.

Normas, confiabilidad y validez

Las calificaciones en los inventarios de personalidad se interpretan, por lo regular, con referencia a un conjunto de normas basadas en las respuestas de grupos seleccionados de personas. Dado que las muestras de estandarización con frecuencia son bastante pequeñas y quizá no representativas de la población (objetivo) a la que se pretende llegar, las normas deben interpretarse con cautela. Además, las calificaciones y normas obtenidas en algunos inventarios de personalidad, sobre todo en los que están conformados por reactivos que tienen un formato de elección forzada, son *ipsativas*. Esto significa que la calificación de una persona en una escala es afectada por sus calificaciones en las escalas restantes. Las calificaciones ipsativas se compensan entre sí, por lo que las calificaciones de una persona en todas las escalas no pueden ser en la misma dirección (alta o baja). Esto vuelve difícil comparar las calificaciones de personas diferentes en una escala o variable en particular.

El hecho de que los factores situacionales regularmente influyen más en las calificaciones de las variables afectivas que en las de las variables cognoscitivas, ocasiona que las medidas de la personalidad sean más inestables que las medidas de capacidad. Junto con las dificultades para definir las características de la personalidad y diseñar medidas aceptables de éstas, la inestabilidad de las mediciones de la personalidad casi siempre da como resultado que esas medidas tengan menor confiabilidad que las calificaciones obtenidas en pruebas de habilidad o de aprovechamiento.

Además de la confiabilidad modesta, los inventarios de personalidad también tienen validez limitada. La simulación y los grupos de respuesta contribuyen a la baja validez de muchos inventarios usados en el diagnóstico y la clasificación clínicos. Otro factor que afecta la validez de los inventarios de personalidad es la susceptibilidad de los usuarios a creer que grupos de reactivos

(escalas) con nombre similar miden la misma variable. Esto puede ocurrir, por ejemplo, cuando las calificaciones de la escala de ansiedad o depresión en un inventario tienen sólo correlaciones modestas con escalas de nombre similar en otro inventario. Por otro lado, una correlación elevada entre calificaciones en las escalas de dos inventarios diferentes puede ser ilusoria, porque el método para responder a las dos escalas es similar, independientemente del contenido.

INVENTARIOS DE SÍNTOMAS Y DE UN SOLO CONSTRUCTO

Si bien es indudable que la gente ha evaluado la personalidad de los otros desde los albores de la historia humana, los principios formales de la evaluación de la personalidad se remontan apenas al inicio del siglo XX. El primer inventario de alguna importancia, la Hoja de Datos Personales, se elaboró durante la Primera Guerra Mundial por R. S. Woodworth para detectar trastornos emocionales entre los reclutas del ejército estadounidense. Este instrumento de una sola calificación consistía en 116 preguntas de sí-no relacionadas con temores anormales, obsesiones, compulsiones, tics, pesadillas y otros sentimientos y conductas. Cuatro reactivos ilustrativos de la Hoja de Datos Personales son:

> ¿Se siente triste y abatido la mayor parte del tiempo?
>
> ¿A menudo se asusta a la mitad de la noche?
>
> ¿Considera que se ha lastimado al tener muchas relaciones con mujeres?
>
> ¿Alguna vez ha perdido la memoria por algún tiempo? (DuBois, 1970, pp. 160-163.)

Otro de los primeros inventarios de personalidad calificados en una sola variable fue el Estudio de Reacción A-S, un instrumento de opción múltiple diseñado por G. W. y F. H. Allport en 1928 para medir la disposición a ser dominante o sumiso en las relaciones sociales cotidianas.

En la actualidad se dispone de muchos inventarios de una sola calificación o un solo constructo. Algunos ejemplos de los constructos psicológicos que los inventarios han sido diseñados para medir son altruismo, ira, ansiedad, depresión, desesperanza, hostilidad, toma de riesgos, autoconcepto, autoestima, búsqueda de sensaciones y estrés. Entre las medidas de un solo constructo más populares se encuentran los inventarios de Beck y varias medidas del autoconcepto y la autoestima.

Inventarios de Beck

Los cuatro instrumentos de este grupo son el Inventario de Ansiedad de Beck, el Inventario de Depresión de Beck, la Escala de Desesperanza de Beck y la Escala de Ideación Suicida de Beck (por A. T. Beck; Psychological Corporation). Los cuatro inventarios han recibido revisiones favorables con respecto a su contenido, administración y calificación (Carlson, 1998; Dowd, 1998; Fernandez, 1998; Hanes, 1998; Stewart, 1998; Waller, 1998a, b). Estos inventarios constan de 20 a 21 reactivos y pueden completarse en 5 a 10 minutos. El Inventario de Depresión de Beck (BDI) y su revisión, BDI-II (Beck y Steer, 1993), son los más populares y, de hecho, se encuentran entre los inventarios de personalidad más exhaustivamente investigados. Los 21 conjuntos de reactivos en el BDI-II, los cuales fueron escritos de acuerdo con las directrices del DSM-IV para el diagnóstico de la depresión, fueron diseñados para evaluar la intensidad de la depresión en personas normales y pacientes psiquiátricos. Los reactivos están compuestos por cuatro afirmaciones arregladas en orden de gravedad creciente con respecto a un síntoma particular de la depresión,

concentrándose en los síntomas presentes durante las dos semanas previas a la evaluación. Es posible determinar calificaciones separadas en las dos subescalas (cognoscitiva-afectiva y somática-desempeño), así como una calificación total. Por lo que atañe a las clasificaciones por calificación total se tiene: de 0 a 9 se clasifica como "normal", de 10 a 18 como "depresión de leve a moderada", de 19 a 29 como "depresión de moderada a severa", y de 30 y más como "depresión extremadamente severa". Las confiabilidades por consistencia interna (coeficiente alfa) de las calificaciones totales son de hasta .92. En el manual y en los resultados de cientos de estudios conducidos con estos instrumentos se presenta evidencia a favor de la validez del BDI y el BDI-II, incluyendo altas correlaciones con las calificaciones clínicas de la depresión.

La Escala de Desesperanza de Beck (BHS), la cual consta de 20 reactivos, tiene un formato similar al Inventario de Beck de la Depresión. Se diseñó para medir tres aspectos importantes de la desesperanza: los sentimientos acerca del futuro, la falta de motivación y las expectativas. Las calificaciones de la BHS tienen una correlación moderada con las de la BDI, pero se considera que el primer instrumento permite una mejor predicción de la intención suicida y la conducta que el último. Las confiabilidades por consistencia interna presentadas en el manual de la BHS de 1988 son razonablemente altas (.82 a .93 en siete grupos de normas). Sin embargo, los coeficientes de confiabilidad test-retest son muy modestos (.69 después de una semana y .66 después de seis semanas). Al revisar la BHS, Dowd (1992) concluyó que es "un instrumento bien elaborado y validado, con una adecuada confiabilidad" (p. 82). La revisión que hizo Owen (1992) de la BHS también fue positiva, aunque menos entusiasta que la de Dowd.

Las otras dos escalas de Beck de formato similar al BDI y la BHS son el Inventario de Ansiedad de Beck BAI) y la Escala de Ideación Suicida de Beck (BSS). Al igual que las otras escalas de Beck, estos nuevos instrumentos fueron diseñados para adultos de 17 a 80 años y pueden encontrarse en inglés y en español. El BAI fue diseñado para medir la gravedad de la ansiedad en adolescentes y adultos y se ha encontrado que discrimina entre grupos de diagnóstico ansiosos y no ansiosos. Los grupos ansiosos incluían a pacientes con agorafobia, trastorno de angustia, fobia social, trastorno obsesivo-compulsivo y ansiedad generalizada. La BSS fue diseñada para evaluar pensamientos y actitudes suicidas y, por ende, para identificar a individuos en riesgo de cometer suicidio. Las confiabilidades por consistencia interna del BAI y el BSI son altas, pero las confiabilidades test-retest son más modestas. Los estudios de la validez clínica de los dos instrumentos se presentan en los manuales del BAI (Beck, 1990) y el BSI (Beck, 1991).

Autoconcepto y autoestima

El *autoconcepto*, el cual consiste en la forma en que una persona se ve a sí misma, depende de las comparaciones que hace la persona de sus características físicas, capacidades y temperamento con las de otros individuos. El autoconcepto incluye también las actitudes, aspiraciones y roles sociales de la persona. Mientras que el autoconcepto se refiere a las ideas o creencias que un individuo tiene acerca de sí mismo, la *autoestima* consiste en la forma en que el yo es evaluado por la persona. La gente puede llegar a evaluarse de manera elevada (alta autoestima) o baja (baja autoestima).

Las clasificaciones Q, las cuales fueron analizadas en el capítulo 16, son medidas del autoconcepto basadas en una técnica de calificación o clasificación. Entre las medidas más antiguas del autoconcepto y la autoestima que todavía se encuentran en el mercado están los Inventarios Coopersmith de Autoestima (Consulting Psychologists Press), la Escala Piers-Harris de Autoconcepto para Niños (Western Psychological Services) y la Escala de Autoconcepto de Tennessee (Western Psychological Services). Otros inventarios populares de autoconcepto y autoestima son Autoestima Académica Conductual (Consulting Psychologists Press), Dimensio-

nes del Autoconcepto (EdITS), Escala de Autoconcepto del Estudiante (American Guidance Service) e Índice de Autoestima (pro.ed).

Inventarios para el diagnóstico de un trastorno específico

Ansiedad, depresión, hostilidad y muchas otras condiciones mencionadas arriba son sintomáticas de varios trastornos psicológicos, y es posible aplicar inventarios para evaluar esos síntomas con propósitos de diagnóstico. Además, se dispone de inventarios diseñados para identificar o diagnosticar un trastorno específico. Existen inventarios para alcoholismo, personalidad antisocial, personalidad limítrofe, agotamiento, trastornos alimentarios, neuroticismo, pánico, agorafobia, personalidad psicopática, fobia social, abuso de sustancias, trauma y otras condiciones psicopatológicas. Muchos de esos inventarios arrojan calificaciones múltiples, pero el énfasis permanece en un solo trastorno o síndrome.

Entre los inventarios que se concentran en un trastorno específico llaman en particular la atención los que tienen que ver con anorexia, bulimia y otros trastornos alimenticios. De esos instrumentos, los más populares e investigados son el Inventario de Alimentación (Psychological Corporation) y el Inventario de Trastornos Alimenticios (Psychological Assessment Resources). La segunda edición del último instrumento, el Inventario de Trastornos Alimenticios-2 (EDI-2), se diseñó para evaluar una amplia gama de rasgos psicológicos de los trastornos alimenticios, como la anorexia nerviosa y la bulimia nerviosa, en pacientes tan jóvenes como de 11 años. Consta de 91 reactivos de elección forzada (64 reactivos originales más 27 adicionales), cada uno de los cuales es calificado por la persona (de 12 años en adelante) en una escala de seis puntos que va de "siempre" a "nunca". Las respuestas son calificadas en ocho subescalas originales (Pulsión por la Delgadez, Ineficacia, Insatisfacción con el Cuerpo, Desconfianza Interpersonal, Bulimia, Perfeccionismo, Madurez, Temor, y Conciencia Interoceptiva) y en tres subescalas provisionales (Regulación de Impulsos, Inseguridad Social y Ascetismo). Sin embargo, las correlaciones positivas significativas entre la mayoría de las escalas muestran que no representan dimensiones independientes. La mayoría de las correlaciones de las escalas de la EDI-2 con las calificaciones en varios inventarios de personalidad y calificaciones de clínicos son modestas pero significativas. En el manual también se describen casos de muestra y algunas investigaciones. A partir de esos datos, puede concluirse de manera tentativa que la EDI-2 es una herramienta de detección clínica y una medida de resultado útil, así como un auxiliar valioso para los juicios clínicos que atañen a pacientes con trastornos alimentarios. Ha recibido revisiones favorables como herramienta clínica para tratar con la anorexia nerviosa, la bulimia y otros trastornos alimentarios (por ejemplo, Ash, 1995; Schinke, 1995).

INVENTARIOS DE CONTENIDO VALIDADO Y CALIFICACIÓN MÚLTIPLE

El primer inventario de ajuste de calificación múltiple, o multifásico, fue el Inventario de Personalidad de Bernreuter (1931). Constaba de 125 reactivos que debían responderse con sí, no o? por estudiantes de preparatoria, universitarios u otros adultos. Al asignar diferentes pesos numéricos a diferentes reactivos, el Bernreuter se calificaba en seis variables: Tendencia Neurótica, Autosuficiencia, Introversión-Extroversión, Dominio-Sumisión, Sociabilidad y Confianza.

Desde 1930 se han publicado muchos otros inventarios de personalidad de calificación múltiple. Los procedimientos estadísticos de análisis factorial y de codificación de criterio complementan el procedimiento lógico-racional de seleccionar reactivos sobre la base de la validez

de contenido. Los diseñadores de ciertos inventarios han aplicado una combinación de dichos procedimientos. Sin embargo, por conveniencia, aquí se describirán los inventarios de personalidad ilustrativos bajo tres encabezados separados: de contenido validado, basados en el análisis factorial y con criterios codificados.

Los reactivos en los *inventarios de contenido validado* eran seleccionados porque al diseñador (o diseñadores) de la prueba le parecía que medían ciertos rasgos o características de personalidad consideradas importantes. Un ejemplo de un antiguo inventario de este tipo es la Escala de Preferencias Personales de Edwards (por A. L. Edwards; Psychological Corporation), el cual se basa en la teoría de personalidad de necesidad-presión de Henry Murray. Debido a que se interesa en el razonamiento y a menudo es guiado por una teoría de la personalidad más que por pruebas empíricas y de estadística, el enfoque de contenido validado en ocasiones ha sido conocido como un método "racional" o "a priori" para la elaboración de instrumentos. Dos ejemplos de inventarios de contenido validado que se basan, al menos hasta cierto grado, en una teoría de la personalidad son el Indicador de Tipos Psicológicos de Myers-Briggs y el Formato de Investigación de Personalidad.

Indicador de Tipos Psicológicos de Myers-Briggs

El Indicador de Tipos Psicológicos de Myers-Briggs (MBTI) (por K. C. Briggs e I. B. Myers; Consulting Psychologists Press) está compuesto por una serie de reactivos de dos opciones concernientes a las preferencias o inclinaciones en los sentimientos y la conducta. Existen cuatro formas (G, F, K y J) que contienen de 126 a 290 reactivos por forma. Basado en la teoría de los tipos de personalidad de Carl Jung, el MBTI se califica en cuatro escalas bipolares: Introversión-Extroversión (I-E), Sensación-Intuición (S-N), Pensamiento-Sentimiento (T-F) y Juicio-Percepción (J-P). Las combinaciones de calificaciones en esas cuatro categorías de dos partes dan como resultado 16 tipos de personalidad posibles. De este modo, un tipo ENFP es una persona cuyos modos predominantes son: Extrovertido, Intuición, Sentimiento y Percepción; mientras que un tipo ISTJ es una persona cuyos modos predominantes son: Introvertido, Sensación, Pensamiento y Juicio. Por desgracia, el hecho de que no se proporcionan medidas de la actitud a la presentación de la prueba puede conducir a errores en el diagnóstico y la detección con el MBTI.

En el manual del MBTI se proporcionan normas de rangos percentilares, basadas en pequeñas muestras de estudiantes de preparatoria y universidad, para las cuatro calificaciones indicadoras (Myers y McCaulley, 1985). Se informa que las confiabilidades de división por mitades de los cuatro indicadores se encuentran entre .70 y .80, y también se describe una serie de estudios de validez a pequeña escala. Aunque muchos psicólogos no ven de manera favorable la conceptualización de la personalidad en términos de tipos, en Consulting Psychologists Press puede encontrarse una colección impresionante de materiales sobre el Indicador de Tipos de Myers-Briggs. Tales materiales incluyen varias guías de interpretación, libros y materiales para talleres. Los perfiles de calificaciones y varios tipos de informes pueden prepararse por medio de una computadora, y también se dispone de otros recursos y servicios para los usuarios.

Formato de Investigación de Personalidad

Basada en gran medida en la teoría de los rasgos de personalidad de Henry Murray y centrada en áreas del funcionamiento normal más que de la psicopatología, el Formato de Investigación de Personalidad (PRF) (por D. N. Jackson, Sigma Assessment Systems) es un conjunto de cinco escalas de verdadero-falso diseñadas para administrarse desde el sexto grado hasta la adultez. Cada una de las 15 escalas en las formas A y B y las 22 escalas de las formas AA, BB y E cons-

tan de 20 reactivos de verdadero-falso. Además de las escalas de contenido, todas las formas se califican en una Escala de Infrecuencia que consta de reactivos que rara vez se marcan. Las formas AA, BB y E también se califican en una Escala de Deseabilidad Social.

El PRF fue estandarizado en 1,000 universitarios y 1,000 universitarias. Los coeficientes de confiabilidad por consistencia interna y test-retest para las calificaciones en las 14 escalas de contenido común a las cinco formas se agrupan en alrededor de .80, pero las confiabilidades de las seis escalas de contenido adicionales en las formas AA, BB y E se encuentran en los .50. Los coeficientes de validez obtenidos al correlacionar las escalas de contenido con las calificaciones de conducta y una forma de calificación de rasgos elaborada especialmente se encuentran en los .50. En el manual se presenta evidencia a favor de la validez convergente y discriminante de la PRF, que utiliza calificaciones de los compañeros y datos de cientos de estudios.

INVENTARIOS SOMETIDOS A ANÁLISIS FACTORIAL

La meta común de los investigadores que aplican *técnicas de análisis factorial* al análisis de la personalidad ha sido aislar un número relativamente pequeño de factores o rasgos de personalidad que puedan explicar las variaciones en las calificaciones de diferentes inventarios y construir luego una medida de cada factor. La primera aplicación publicada del análisis factorial al estudio de la personalidad fue realizada por Webb (1915), quien formó grupos de estudiantes varones para calificar 40 cualidades que ellos consideraran tenían "una fuerza general y fundamental sobre la personalidad total". El desarrollo subsecuente de las técnicas de análisis factorial durante las décadas de 1930 y 1940 llevó a la elaboración de inventarios multifactoriales de personalidad por L. L. Thurstone, J. P. Guilford, R. B. Cattell, H. Eysenck y otros psicólogos. Algunos ejemplos de esos primeros inventarios basados en factores son el Inventario de Factores STDCR, el Programa de Temperamento de Thurstone y el Estudio de Temperamento de Guilford-Zimmerman.

Cuestionario de 16 Factores de la Personalidad

La serie más amplia de inventarios basados en factores para evaluar la personalidad en niños y adultos fue diseñada por R. B. Cattell y publicada por el Instituto para la Personalidad y las Pruebas de Habilidad. Cattell comenzó su investigación de la personalidad con una lista de alrededor de 18,000 adjetivos descriptivos de la personalidad que Allport y Odbert (1936) habían recopilado de los diccionarios. Al combinar los términos que tenían significados similares, la lista fue reducida primero a 4,500 rasgos "reales" y luego a 171 nombres de rasgos; un análisis factorial subsecuente de las calificaciones obtenidas en esas dimensiones de rasgo produjo 31 rasgos superficiales y 12 rasgos fuente de personalidad. Cattell desarrolló una serie de medidas de esos rasgos y de otros cuatro que aisló en su trabajo posterior, pero su producto principal fue el Cuestionario de 16 Factores de la Personalidad (16 PF).

La quinta edición del 16 PF consta de 185 reactivos de tres opciones, incluyendo de 10 a 15 reactivos por cada una de las 16 escalas de factores primarios (Russell y Karol, 1994). Los reactivos en la quinta edición del 16 PF reflejan el uso moderno del lenguaje y se les analizó para detectar ambigüedad, así como sesgos de género, raza y cultura. La legibilidad global de este inventario se encuentra al nivel del quinto grado y el tiempo total de completamiento del examen es de 35 a 50 minutos. Además de los 16 factores primarios, el 16 PF puede ser calificado, a mano o por computadora, en tres índices de validez y cinco calificaciones globales (factores de segundo orden). Estos índices (Manejo de Impresiones, Infrecuencia y Conformidad) proporcionan una verificación preliminar sobre la validez de las respuestas.

El resumen o reporte de calificaciones del 16 PF generado por computadora contiene interpretaciones narrativas de los índices de validez, las calificaciones globales, funcionamiento cognoscitivo y perceptual, estilo interpersonal, relaciones íntimas, consideraciones ocupacionales, dinámicas de personalidad y aspectos terapéuticos y de orientación. Los datos normativos para el 16 PF se basan en el censo estadounidense de 1990 y se dispone de normas combinadas de género. Además de la mejoría en la elaboración y las normas, las escalas de la quinta edición del 16 PF tienen confiabilidades mayores que las de ediciones previas. Las confiabilidades por consistencia interna van de .64 a .85, con un promedio de .74; las confiabilidades test-retest promedian alrededor de .80 luego de un intervalo de dos semanas y .70 luego de dos meses.

A principios del 2002 se dispuso de normas actualizadas para la quinta edición del cuestionario 16 PF. Esas normas se basan en las respuestas de adultos en una muestra estratificada para igualar las cifras del censo del 2000 de la población general de Estados Unidos.

Inventario de la personalidad adulta

Relacionado con el 16 PF está el Inventario de Personalidad para Adultos (API) (por S. F. Krug; MetriTech; también vea Krug, 1999), el cual, junto con el Inventario Multifásico de Personalidad de Minnesota (MMPI) y el Inventario de Evaluación de la Personalidad (PAI), recibió el mayor número de citas de investigación de la personalidad durante los pasados seis años. El API es un inventario de autorreporte de 324 reactivos para evaluar la personalidad en adultos normales y puede ser calificado en 21 escalas de contenido y cuatro escalas de validez. Las escalas de contenido constan de siete características de personalidad (extrovertido, ajustado, realista, independiente, disciplinado, creativo y emprendedor), ocho estilos interpersonales (preocupado, adaptado, aislado, sumiso, despreocupado, no conformista, sociable y asertivo) y seis factores de estilo de vida o de carrera (práctico, científico, estético, social, competitivo y estructurado).

El API fue estandarizado en 1,000 adultos y se dispone de normas separadas para hombres y mujeres. Sin embargo, se ha criticado a las normas por ser poco representativas (D'Amato, 1995). La información sobre la confiabilidad y validez de constructo que se presenta en el manual es muy limitada. Los coeficientes de confiabilidad por consistencia interna y test-retest promedian alrededor de .75. A pesar de esas desventajas, el API ha sido utilizado en varios contextos de orientación y personal. La disponibilidad de un software de computadora para el Perfil de Carreras ha contribuido a la popularidad de este inventario entre los profesionales e investigadores.

Cuestionario de Personalidad de Eysenck

El Cuestionario de Personalidad de Eysenck (EPQ) (por H. Eysenck; EdITS), una revisión del Inventario de Personalidad de Eysenck y del Inventario de Personalidad Junior de Eysenck, representa un concepto más moderado de la personalidad que el que se refleja en los inventarios de Cattell. Dos inventarios anteriores diseñados por Eysenck, el Inventario de Personalidad Maudsley y el Inventario de Personalidad de Eysenck, se calificaban en las dimensiones de neuroticismo (N) y extroversión (contra introversión) (E) que surgieron de su investigación analítico-factorial. Al elaborar el EPQ se agregaron una medida del psicoticismo (P) y una escala de mentiras (L).

El EPQ tiene un rango de edad amplio (de los siete años a la adultez) y sólo se necesitan de 10 a 15 minutos para completarlo. Las confiabilidades test-retest de las escalas N, E, P y L del EPQ van de .78 a .80 luego de un intervalo de un mes; los coeficientes de consistencia interna están entre .70 y .80. Las normas de las dos formas (A y B), basadas en universitarios y adultos estadounidenses, son apropiadas para individuos de 16 años en adelante. Las normas en el EPQ Junior fueron obtenidas de muestras de niños de 7 a 15 años de edad. El EPQ y sus predecesores han si-

do muy utilizados en la investigación de la personalidad, aunque con menos frecuencia en los contextos clínicos y otros contextos aplicados. Eysenck (1965, 1981) utilizó calificaciones de los factores E y N en particular para predecir cómo reaccionaría la gente en ciertas situaciones experimentales. También relacionó los patrones de personalidad con el tipo corporal.

Perspectiva sobre el análisis factorial

Muchos otros inventarios de personalidad han sido elaborados usando los métodos del análisis factorial. Sin embargo, independientemente de la complejidad matemática de esos métodos, la mayoría de los psicómetras no cree que el análisis factorial identifique dimensiones "verdaderas" o "reales" de la personalidad. Lo que revela son consistencias internas y diferencias entre los reactivos de la prueba y las escalas, aclarando así las relaciones entre los constructos o variables de personalidad.

Debido a que la validez relacionada con el criterio de los inventarios de personalidad elaborados mediante análisis factorial tiende a ser baja o desconocida, esos inventarios son, por lo general, menos útiles que los instrumentos de contenido validado y con criterios codificados para formular predicciones conductuales y tomar decisiones en la clínica y otros contextos psicológicos aplicados. No obstante, muchos psicólogos encuentran atractiva la aplicación del análisis factorial a la elaboración de inventarios de personalidad y a la investigación básica sobre la naturaleza de la personalidad humana. Existe un acuerdo bastante general de que muchos inventarios de personalidad miden al menos los factores de extroversión-introversión y neuroticismo (emocionalidad) descritos por Eysenck. Además, la evidencia a favor del modelo de cinco factores de personalidad es impresionante. Goldberg (1980) designó esos cinco factores como extroversión o surgencia, agradabilidad, escrupulosidad, estabilidad emocional y cultura; Costa y McRae (1986) definieron los cinco factores de personalidad, los cuales parecen ser muy consistentes entre varios grupos de personas y situaciones distintas, de la siguiente manera:

> *Neuroticismo*: Preocupado contra tranquilo, inseguro contra seguro, autocompasivo contra autosatisfecho.
>
> *Extroversión*: Sociable contra recluido, amante de la diversión contra solemne, afectuoso contra reservado.
>
> *Apertura*: Imaginativo contra realista, preferencia por la variedad contra preferencia por la rutina, independiente contra conformista.
>
> *Agradabilidad*: Bondadoso contra despiadado, confiado contra suspicaz, útil contra poco cooperativo.
>
> *Escrupulosidad*: Bien organizado contra desorganizado, cuidadoso contra descuidado, autodisciplinado contra carente de voluntad.

Inventario NEO de Personalidad

El Inventario NEO de Personalidad, revisado (NEO-PI-R) y una versión abreviada, el Inventario NEO de Cinco Factores (NEO-FFI) (por P. T. Costa, Jr., y R. R. McCrae; Psychological Assessment Resources), se basan en el modelo de cinco factores descrito líneas arriba. Cada una de las dos formas (R y S) del NEO-PI-R consta de 240 reactivos que deben ser calificados en una escala de cinco puntos y requieren aproximadamente 30 minutos para completarse. El NEO-FFI consta de 60 reactivos y sólo se lleva de 10 a 15 minutos completarlo. Tanto el NEO-PI-R como el NEO-FFI se califican en los tres *dominios* (factores) N-E-O: Neuroticismo (N), Extroversión (E) y

Apertura a la Experiencia (O), además de Agradabilidad (A) y Escrupulosidad (C). Cada uno de estos cinco dominios se subdivide además en seis *facetas* calificables de la siguiente manera:

Neuroticismo: Ansiedad, hostilidad, depresión, conciencia de sí mismo, impulsividad, vulnerabilidad.

Extroversión: Calidez, carácter gregario, asertividad, actividad, búsqueda de sensaciones, emociones positivas.

Apertura a la experiencia: Fantasía, estética, sentimiento, acciones, ideas, valores.

Agradabilidad: Confianza, modestia, condescendencia, altruismo, sinceridad, idealismo.

Escrupulosidad: Competencia, autodisciplina, esfuerzo por el logro, cumplimiento de los deberes, orden, deliberación.

Los coeficientes de confiabilidad por consistencia interna de las calificaciones en las escalas de dominio van de .86 a .95 para el NEO-PI-R y de .68 a .86 para el NEO-FFI. Los coeficientes de consistencia interna para las escalas de facetas del NEO-PI-R van de .56 a .90. Las confiabilidades test-retest calculadas luego de un periodo de seis meses van de .86 a .91 para las escalas de dominio y de .56 a .90 para las escalas de facetas. La evidencia a favor de la validez de esos inventarios es algo escasa, pero en el manual se informa de correlaciones con otros inventarios de personalidad, calificaciones de expertos y calificaciones de pruebas de frases incompletas.

INVENTARIO MULTIFÁSICO DE PERSONALIDAD DE MINNESOTA

Al igual que el Inventario de Intereses de Strong, los *inventarios de personalidad con codificación de criterios* están compuestos por reactivos o escalas que diferencian entre dos o más grupos de criterios. Uno de los primeros instrumentos de este tipo fue el Estudio de Reacción A-S, el cual constaba de reactivos que diferenciaban entre grupos de personas que habían sido calificadas por sus compañeros como dominantes o sumisas. Sin embargo, el inventario de personalidad con codificación de criterios más famoso es el Inventario Multifásico de Personalidad de Minnesota (MMPI).

Descripción del MMPI

La primera edición del MMPI fue diseñada a principios de la década de 1940 por S. R. Hathaway y J. C. McKinley para evaluar características de personalidad que indican una anormalidad psicológica en los adultos. Aunque en gran medida ha sido reemplazado por una segunda edición (MMPI-2), el diseño, la validación y el uso del MMPI original proporcionaron un antecedente y directrices para otros inventarios de personalidad desarrollados mediante el enfoque empírico.

Las 550 afirmaciones del MMPI, las cuales se responden con sí, no o no podría decirlo, se interesan en las actitudes, emociones, perturbaciones motrices, síntomas psicosomáticos y otros sentimientos y conductas reportadas que son indicadores de problemas psiquiátricos. Cada una de las nueve escalas sobre las cuales se califica el MMPI consta de reactivos que fueron respondidos de manera diferente por pacientes psiquiátricos en un grupo especificado de diagnóstico y por un grupo control de gente normal. En la tabla 17.1 se describen las nueve escalas clínicas,

junto con la escala Si (introversión social) y las cuatro escalas de validez (*?*, *L*, *F*, *K*). Muchas escalas especiales (por ejemplo, proclividad a los accidentes, ansiedad, fortaleza del Yo, originalidad) fueron desarrolladas a partir del grupo de reactivos del MMPI durante el curso de miles de investigaciones conducidas a lo largo de medio siglo.

TABLA 17.1 Descripción de las escalas de validez y clínicas del MMPI original

Escalas de validez (actitud hacia la presentación de la prueba)

? (*No podría decirlo*) Número de reactivos que se dejan sin responder.

L (*Mentira*) Quince reactivos de autorreporte demasiado bueno, como "Sonrío a todos los que encuentro". (Respondido como Verdadero.)

F (*Frecuencia o infrecuencia*) Sesenta y cuatro reactivos respondidos en la dirección calificada por 10% o menos de las personas normales, como "Hay una conspiración internacional en mi contra". (Verdadero)

K (*Corrección*) Treinta reactivos que reflejan una posición defensiva al admitir problemas, como "Me siento mal cuando otros me critican". (Falso)

Escalas clínicas

1 o *Hs* (*Hipocondriasis*) Treinta y tres reactivos derivados de pacientes que muestran una preocupación anormal por las funciones corporales, como "Tengo dolores en el pecho varias veces a la semana". (Verdadero)

2 o *D* (*Depresión*) Sesenta reactivos derivados de pacientes que muestran un pesimismo extremo, sentimientos de desesperanza y aletargamiento del pensamiento y la acción, como "Por lo regular siento que la vida es interesante y valiosa". (Falso)

3 o *Hy* (*Histeria de conversión*) Sesenta reactivos de pacientes neuróticos que utilizan síntomas físicos o mentales como una forma de evitación inconsciente de los conflictos y las responsabilidades, como "Con frecuencia mi corazón late tan fuerte que puedo sentirlo". (Verdadero)

4 o *Pd* (*Desviación psicopática*) Cincuenta reactivos de pacientes que muestran un descuido repetido y flagrante por las costumbres sociales, una superficialidad emocional y una incapacidad para aprender de las experiencias de castigo, como "Mis actividades e intereses a menudo son criticados por los demás". (Verdadero)

5 o *Mf* (*Masculinidad-Feminidad*) Sesenta reactivos de pacientes que muestran homoerotismo y reactivos que diferencian entre hombres y mujeres, como "Me gusta arreglar las flores". (Verdadero, calificado para feminidad)

6 o *Pa* (*Paranoia*) Cuarenta reactivos de pacientes que muestran suspicacia anormal y delirios de grandeza o persecución, como "Hay personas malvadas que tratan de influir mi mente". (Verdadero)

7 o *Pt* (*Psicastenia*) Cuarenta y ocho reactivos basados en pacientes neuróticos que muestran obsesiones, compulsiones, temores anormales, culpa e indecisión, como "Guardo todo lo que compro incluso después de que no lo uso". (Verdadero)

8 o *Sc* (*Esquizofrenia*) Sesenta y ocho reactivos de pacientes que muestran conducta o pensamientos extraños o inusuales, quienes se aíslan con frecuencia y experimentan delirios o alucinaciones, como "Las cosas a mi alrededor no parecen reales" (Verdadero); y "Me hace sentir incómodo tener gente cerca de mi". (Verdadero)

9 o *Ma* (*Hipomanía*) Cuarenta y seis reactivos de pacientes caracterizados por excitación emocional, actividad excesiva y vuelo de ideas, como "En ocasiones me siento muy alto o muy bajo sin razón aparente". (Verdadero)

0 o *Si* (*Introversión social*) Setenta reactivos de personas que muestran timidez, poco interés en la gente e inseguridad, como "Paso el tiempo de mi vida en fiestas".

Fuente: Tomado de Sundberg (1977). Los reactivos citados son reactivos simulados del MMPI. Los nombres y abreviaturas de las escalas del MMPI son del Inventario Multifásico de Personalidad de Minnesota.

Antes de intentar interpretar las calificaciones en las escalas clínicas o especiales del MMPI, deben inspeccionarse las calificaciones en las cuatro escalas de validez. La primera de éstas, la puntuación cruda de la pregunta (*?*), es el número total de reactivos que el examinado respondió con "no podría decirlo" o que no respondió. Una calificación alta de la pregunta se interpreta como una posición defensiva al responder. La puntuación cruda a Mentira (*L* o fingir para verse bien) es el número de reactivos respondidos de tal manera que uno se coloca bajo una luz más favorable, mientras que la calificación a infrecuencia (*F* o fingir para verse mal) es el número de reactivos que se responden de tal manera que uno se coloca bajo una luz menos favorable. La gente a menudo miente para verse bien a fin de obtener algo placentero, mientras que miente para verse mal a fin de evitar algo desagradable, como ir a prisión, al servicio militar u otras consecuencias desagradables.

La calificación *K*, una fracción de la cual se aplica como factor de corrección a las puntuaciones crudas en las escalas clínicas 1, 4, 7, 8 y 9, es una medida de la crítica o generosidad excesivas al evaluarse uno mismo. Quienes califican alto en la escala *K* tienden a negar las insuficiencias y deficiencias personales en autocontrol; quienes califican bajo están dispuestos a decir cosas que son socialmente indeseables acerca de sí mismos.

El MMPI revisado

En la década de 1980 se realizó una revisión del MMPI por las siguientes razones: proporcionar normas nuevas y actualizadas; ampliar la base de reactivos con contenido no representado en la versión original; revisar y replantear el lenguaje de algunos de los reactivos existentes que eran anticuados, inconvenientes o sexistas, y proporcionar formas separadas del inventario para adultos y adolescentes. Los 550 reactivos del MMPI original fueron conservados en las versiones revisadas para adultos y adolescentes, pero 14% de ellos fue cambiado por contener un lenguaje anticuado o expresiones inconvenientes. Se omitieron las palabras o frases que eran más características de la década de 1940 (tranvía, polvo para dormir, dejar caer el pañuelo, etc.) y se hicieron otras modificaciones para actualizar las afirmaciones (por ejemplo, "Me gusta tomar un baño" se convirtió en "Me gusta tomar un baño o una ducha"). Al igual que en la forma original, en el MMPI revisado los reactivos se escribieron al nivel de sexto grado. La versión para adultos (MMPI-2) contenía 154 reactivos experimentales nuevos diseñados para evaluar ciertas áreas de la psicopatología (como trastornos alimenticios, personalidad tipo A y abuso de drogas) que no estaban bien representadas en el MMPI original. La versión para adolescentes (MMPI-A) contenía 104 reactivos nuevos que se referían de manera específica a problemas de los adolescentes. Además, se corrigió la tendencia a que adolescentes normales que pasan por un estado temporal de confusión califiquen como los psicópatas adultos en el MMPI original.

Diseñado para adecuarse a usos tanto no clínicos como clínicos, el MMPI-2 consta de 567 preguntas de verdadero-falso escritas a nivel de octavo grado y se lleva alrededor de 90 minutos para responderse. Las cuatro escalas de validez y las diez escalas clínicas básicas se califican a partir de los primeros 370 reactivos, mientras que las escalas complementarias de contenido e investigación se califican del reactivo 371 al 567 (vea la tabla 17.2).

El MMPI-2 se califica en las mismas escalas clínicas que el MMPI, pero las calificaciones *T* para ocho escalas clínicas y otras escalas (de contenido) se han uniformado. Las calificaciones *T* uniformes se determinaron porque, debido a las diferencias en las distribuciones de calificación, las calificaciones *T* tradicionales en las diferentes escalas no eran estrictamente comparables. Esas diferencias fueron eliminadas al uniformar las calificaciones *T* las cuales, a diferencia de las calificaciones *T* normalizadas, conservan la forma general de las distribuciones de calificación cruda.

TABLA 17.2 Escalas del MMPI-2

ESCALAS DE VALIDEZ

VRIN	Inconsistencia de la respuesta variable
TRIN	Inconsistencia de la respuesta verdadera
F	Infrecuencia
F_B	Regresar a F
F_P	F-Psicopatología
L	Mentira
K	Posición defensiva
S	Autopresentación superlativa
?	No podría decirlo

Subescalas de autopresentación superlativa

S_1	Creencias en la bondad humana
S_2	Serenidad
S_3	Satisfacción con la vida
S_4	Paciencia/Negación de la irritabilidad
S_5	Negación de los defectos morales

ESCALAS CLÍNICAS

1 *Hs*	Hipocondriasis
2 *D*	Depresión
3 *Hy*	Histeria de conversión
4 *Pd*	Desviación psicopática
5 *Mf*	Masculinidad-Feminidad
6 *Pa*	Paranoia
7 *Pt*	Psicastenia
8 *Sc*	Esquizofrenia
9 *Ma*	Hipomanía
0 *Si*	Introversión social

SUBESCALAS CLÍNICAS

Subescalas Harris-Lingoes

*D*1	Depresión subjetiva
*D*2	Retardo psicomotriz
*D*3	Mal funcionamiento físico
*D*4	Torpeza mental
*D*5	Meditación
*Hy*1	Negación de la ansiedad social
*Hy*2	Necesidad de afecto
*Hy*3	Lasitud-malestar
*Hy*4	Quejas somáticas
*Hy*5	Inhibición de la agresión
*Pd*1	Discordia familiar
*Pd*2	Problemas de autoridad
*Pd*3	Imperturbabilidad social
*Pd*4	Alienación social
*Pd*5	Autoalienación
*Pa*1	Ideas persecutorias
*Pa*2	Viveza
*Pa*3	Ingenuidad
*Sc*1	Alienación social
*Sc*2	Alienación emocional
*Sc*3	Falta de dominio del ego-cognoscitivo
*Sc*4	Falta de dominio del ego-conativo
*Sc*5	Falta de dominio del ego-inhibición defectuosa
*Sc*6	Experiencias sensoriales extrañas
*Ma*1	Amoralidad
*Ma*2	Aceleración psicomotriz
*Ma*3	Imperturbabilidad
*Ma*4	Inflación del Yo

Subescalas de introversión social

*Si*1	Timidez/Conciencia de sí mismo
*Si*2	Evitación social
*Si*3	Alienación del yo y los otros

ESCALAS DE CONTENIDO

ANX	Ansiedad
FRS	Temores
OBS	Obsesividad
DEP	Depresión
HEA	Preocupaciones por la salud
BIZ	Ideas extravagantes
ANG	Ira
CYN	Cinismo
ASP	Prácticas antisociales
TPA	Tipo A
LSE	Baja autoestima
SOD	Incomodidad social
FAM	Problemas familiares
WRK	Interferencia en el trabajo
TRT	Indicadores negativos del tratamiento

ESCALAS DE LOS COMPONENTES DEL CONTENIDO

Subescalas de temores

FRS1	Pusilanimidad generalizada
FRS2	Temores múltiples

Subescalas de depresión

DEP1	Carencia de pulsión
DEP2	Disforia
DEP3	Menosprecio por uno mismo
DEP4	Ideación suicida

Subescalas de preocupaciones por la salud

HEA1	Síntomas gastrointestinales
HEA2	Síntomas neurológicos
HEA3	Preocupaciones generales por la salud

(continúa)

TABLA 17.2 Continuación

Subescalas de ideas extravagantes
BIZ1 Sintomatología psicótica
BIZ2 Características esquizotípicas

Subescalas de ira
ANG1 Conducta explosiva
ANG2 Irritabilidad

Subescalas de cinismo
CYN1 Creencias misantrópicas
CYN2 Suspicacia interpersonal

Subescalas de prácticas antisociales
ASP1 Actitudes antisociales
ASP2 Conducta antisocial

Subescalas tipo A
TPA1 Impaciencia
TPA2 Pulsión competitiva

Subescalas de baja autoestima
LSE1 Duda de uno mismo
LSE2 Sumisión

Incomodidad social
SOD1 Introversión
SOD2 Timidez

Problemas familiares
FAM1 Discordia familiar
FAM2 Alienación familiar

Indicadores negativos del tratamiento
TRT1 Baja motivación
TRT2 Incapacidad para revelar

ESCALAS COMPLEMENTARIAS

Cinco escalas de psicopatología de la personalidad (PSY-5)
AGGR Agresividad
PSYC Psicoticismo
DISC Sin apremio
NEGE Emocionalidad negativa/Neuroticismo
INTR Introversión/Baja emocionalidad positiva

A Ansiedad
R Represión
Es Fortaleza del Yo
Do Dominio
Re Responsabilidad social
Mt Mal ajuste a la universidad
PK Trastorno de estrés postraumático-Keane
MDS Escala de aflicción matrimonial
Ho Hostilidad
O-H Hostilidad controlada en exceso
MAC-R MacAndrew-Revisada
AAS Admisión de la adicción
APS Adicción potencial
GM Rol de género masculino
GF Rol de género femenino

ÍNDICES ESPECIALES
Códigos Welsh (basados en las normas del MMPI-2 y el MMPI)
Índice de disimulación F-K
Porcentaje verdadero y porcentaje falso
Elevación del perfil promedio
Sistema de Clasificación Megargee para Transgresores Criminales
Clasificación P-A-I-N

Fuente: Tomado del *Inventario Multifásico de Personalidad de Minnesota-2 (MMPI-2) Manual de aplicación, calificación e interpretación. Edición revisada.*

Para proporcionar una muestra más representativa de los estadounidenses adultos que sus predecesores, el MMPI-2 fue estandarizado en 1,138 varones y 1,462 mujeres (de 18 a 90 años de edad) residentes de Estados Unidos. La muestra de estandarización fue seleccionada de acuerdo con los datos del censo de 1980, sobre la base de la distribución geográfica, la composición étnica y racial, los niveles de edad y educación y el estado civil. Los datos de confiabilidad presentados en el manual del MMPI-2 (Hathaway y McKinley, 1989) se basan en muestras relativamente pequeñas (82 hombres y 111 mujeres); los coeficientes test-retest para las calificaciones en las escalas básicas van de .58 a .92. Algunos de los coeficientes de confiabilidad bajos, junto con los considerables errores estándar de medición, indican que las diferencias en las calificaciones en las diversas escalas deben interpretarse con cautela.

Interpretación de los perfiles del MMPI-2

La figura 17.1 es un perfil de las calificaciones obtenidas en el MMPI-2 por un hombre de negocios de 60 años de edad descrito en el informe 17.1. Aunque un perfil general alto en las escalas clínicas sugiere problemas psicológicos graves, una calificación *T* alta en una determinada escala clínica no por fuerza es indicativa del trastorno con el que se etiqueta a la escala. Por ésta y otras razones, las escalas clínicas son identificadas por sus designaciones numéricas. En lugar de basarse en una sola calificación, un diagnóstico psiquiátrico o análisis de la personalidad se elabora sobre la base del patrón mostrado por todo el grupo de calificaciones.

Se han elaborado varios sistemas para codificar los perfiles de calificación en el MMPI, de los cuales los más populares son los de Hathaway y Welsh. El proceso de codificación comienza con el ordenamiento de las designaciones numéricas de las nueve escalas clínicas y la escala de Introversión Social (Escala 0), de izquierda a derecha, en orden descendente de sus calificaciones *T*. La realización de este proceso de ordenamiento para las calificaciones perfiladas en la figura 17.1 da por resultado 1267039845. Ambos sistemas de codificación de perfiles de Hathaway y Welsh requieren que se coloque un apóstrofo (´) después del número de la última escala que tiene una calificación *T* de 70 o más, y un guión (-) después del número de la última escala que tiene una calificación *T* de 60 o más. Las designaciones numéricas de las escalas

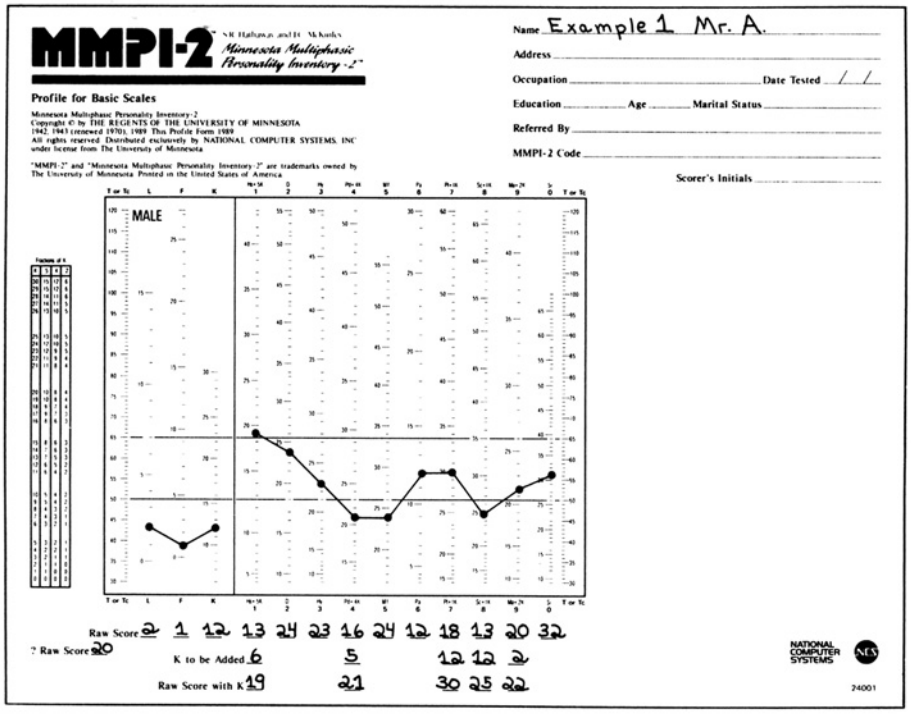


FIGURA 17.1 Perfil de muestra de las calificaciones en el MMPI-2. Vea el informe 17.1.

(Tomado del *Inventario Multifásico de Personalidad de Minnesota-2* (*MMPI-2*) *Perfil para las escalas básicas*. Derechos reservados © 1989 por los Miembros del Directorio de la Universidad de Minnesota. Reproducido con autorización del editor. Todos los derechos reservados. "MMPI-2" e "Inventario Multifásico de Personalidad de Minnesota-2" son marcas registradas propiedad de la Universidad de Minnesota.)

Informe 17.1 Informe interpretativo de las calificaciones dadas en el perfil del MMPI-2 en la figura 17.1

El señor A fue visto en un servicio médico de consulta externa porque se quejaba de una serie de dolores y molestias abdominales. Es un hombre de negocios de sesenta años, blanco, casado, con dos años de universidad. Se pudo encontrar poca evidencia de una base orgánica para sus quejas, y fue canalizado para recibir evaluación psicológica.

En la figura 17.1 se muestra el perfil que obtuvo en el MMPI-2; el código es 12-670 39/845/LK:F. Todos los indicadores tradicionales de validez están por debajo de la media y sugieren que fue muy cooperativo en la prueba. No hay evidencia de posición defensiva o de intención de distorsionar la presentación de sí mismo en el inventario. Sus calificaciones L y K cayeron en los rangos que elevan la posibilidad de que estuviera simulando deliberadamente un mal ajuste, pero su calificación en la escala F no indica que esto sea verdad. Los correlatos de esos indicadores de validez sugieren que este hombre es abierto y convencional, es probable que exponga sus problemas, pero no se encuentra inmerso en una crisis emocional grave.

En la escala 1, su calificación clínica más alta, obtuvo una calificación T de 66. Una calificación en el rango elevado de esta escala sugiere que es más bien egocéntrico y exigente, pesimista y fatalista en su visión del futuro, y propenso a sobreactuar ante cualquier problema real. Es probable que el señor A tenga en el futuro numerosas molestias físicas que se manifestarán en diferentes partes del cuerpo.

Su segunda calificación más alta es en la escala 2 y cae en el rango moderado. Esta calificación también sugiere que es pesimista y desalentado acerca del futuro. Está insatisfecho consigo mismo o con el mundo, se preocupa y es malhumorado. Aunque de temperamento introvertido, es un individuo responsable y modesto.

Otras tres calificaciones caen dentro del rango moderado; las escalas 6, 7 y 0. Estas calificaciones también caracterizan al señor A como responsable, trabajador y reservado.

Los individuos con los perfiles 12/21 muestran una reacción exagerada a los trastornos físicos; son propensos a la fatiga y a menudo son tímidos, irritables, aislados y deprimidos. El dolor visceral, la preocupación excesiva por las funciones del cuerpo y la carencia de introspección son rasgos destacados.

El análisis escala por escala del perfil de esta persona pone de relieve algunas tendencias hipocondríacas y depresivas en un hombre introvertido, malhumorado y trabajador. Las características del código tipo están presentes pero sólo en un grado moderado, como era de esperar para las elevaciones del perfil de esta magnitud.

Estas caracterizaciones surgen en la información de los antecedentes del señor A. Se casó a la edad de 25 años con su esposa actual; no han tenido dificultades matrimoniales. Sin embargo, la señora A dejó recientemente su trabajo, lo que dio como resultado una disminución en el ingreso familiar. Tienen un hijo de 25 años que vive fuera de la casa.

El señor A ha consultado a su médico familiar con mucha frecuencia en el último año y en los últimos meses ha hecho tres visitas a la clínica de consulta externa de la Administración de Veteranos. Además de sus problemas abdominales, el señor A ha tenido problemas de sueño, quejas de fatiga crónica, pérdida del interés en el sexo y temores recurrentes de muerte. También presenta una pérdida considerable de peso y ha tenido dificultades para concentrarse en el trabajo. Los sedantes no han sido de ayuda. La impresión diagnóstica actual es que el señor A sufre de distimia (depresión moderada) con rasgos hipocondríacos.

Fuente: Tomado de *The Minnesota Multiphasic Report™, Adult Clinical System, Revised.*

que tienen valores de calificación *T* dentro de un punto entre sí se subrayan, y las calificaciones en las escalas *L*, *F* y *K* se colocan después del código del perfil. El código Welsh completo para el perfil en la figura 17.1 es 12-670 39/845/LK:F.

Al interpretar un perfil del MMPI o del MMPI-2 se brinda especial atención a las escalas que tienen calificaciones *T* elevadas (por encima de 65 o 70). Se considera que la escala 2 es una medida de depresión, y que la escala 7 es una medida de ansiedad, tensión o alerta ante un peligro des-

conocido. Como la depresión y la ansiedad son síntomas comunes de trastorno mental, los pacientes psiquiátricos tienen, por lo general, calificaciones elevadas en la escala 2, la escala 7 o en ambas. Calificaciones altas en las escalas 2 y 7 señalan una combinación de ansiedad y depresión. Otros patrones de altas calificaciones indican otros síntomas. Por ejemplo, calificaciones altas en las escalas 4 y 9 sugieren impulsividad, baja tolerancia a la frustración, rebeldía y agresión hostil. Las calificaciones elevadas en las escalas 6 y 8 indican aislamiento, apatía y delirios paranoides.

En las escalas clínicas del MMPI algunos términos especiales han sido asociados con ciertos patrones de calificaciones altas. Las escalas 1, 2 y 3 se conocen como la *tríada neurótica* porque las personas con altas calificaciones en esas escalas a menudo tienen problemas psiconeuróticos. Cuando las calificaciones *T* en esas tres escalas están por encima de 70, pero la escala 2 es más baja que las escalas 1 y 3, la configuración se conoce como una *conversión V* y se asocia con un diagnóstico de histeria de conversión. En el otro extremo del perfil, las escalas 6, 7, 8 y 9 se conocen como la *tétrada psicótica* debido a su asociación con los problemas psicóticos. Una configuración en la cual las calificaciones *T* en las escalas 6 a 9 estén por encima de 70, pero las calificaciones *T* en las escalas 7 y 9 sean menores que en las escalas 6 y 8, se conoce como un *valle paranoide* y sugiere un diagnóstico de esquizofrenia paranoide.

Interpretación de pruebas con base en una computadora

Con base en un conjunto de reglas establecidas para el análisis de configuraciones o patrones de calificaciones del MMPI y el MMPI-2, se ha desarrollado una serie de programas de cómputo interpretativos. A pesar de la aparente experiencia y credibilidad de muchas interpretaciones basadas en la computadora, debe tenerse cuidado de no mecanizar en exceso el proceso de interpretación de perfiles. Al igual que las personas, las computadoras cometen errores, pero podemos estar inclinados a confiar más en las últimas que en las primeras.

Los primeros programas para la interpretación de pruebas basados en la computadora (CBTI) se desarrollaron a principios de la década de 1960 en la Clínica Mayo, el Instituto de la Vida de Hartford y la Universidad de Alabama (Glueck y Reznikoff, 1965; Rome *et al*., 1962; Swenson y Pearson, 1964). Estos programas fueron diseñados para calificar, perfilar e interpretar las respuestas al MMPI que habían sido registradas a mano sobre hojas de respuesta de registro óptico. Más tarde, Fowler (1966-1976) desarrolló un programa más complejo para la interpretación automatizada del MMPI. Hacia mediados de la década de 1970, Johnson y Williams (1975) habían diseñado un sistema de interpretación de pruebas a gran escala basado en la computadora para los pacientes del Hospital de la Administración de Veteranos en la ciudad de Salt Lake. En 1977 se tuvo acceso a un sistema en línea mejorado de interpretación del MMPI para usar con microcomputadoras.

En la actualidad se dispone comercialmente de cientos de programas y servicios CBTI, incluyendo programas que califican e interpretan los resultados de pruebas de habilidades cognoscitivas, funcionamiento neuropsicológico y personalidad. Por ejemplo, el informe 17.1 fue generado por un programa desarrollado para la interpretación del MMPI-2. Muchas compañías proporcionan servicios de interpretación e informe computarizado de las pruebas, así como hardware y software para la examinación basada en computadora.

Los programas que generan informes CBTI representan ya sea una destilación de la experiencia clínica o un conglomerado de las relaciones estadísticas entre lo que la gente reporta en un inventario de personalidad u otro instrumento psicométrico y la forma en que se comporta en realidad. En el enfoque clínico, el programa de diagnóstico imita, en esencia, los juicios clínicos de personas expertas en elaborar diagnósticos psicológicos. En el enfoque estadístico, o actuarial, las

afirmaciones interpretativas reportadas se basan en diferencias significativas entre las respuestas de dos grupos contrastantes de personas, aquellos que fueron asignados de manera independiente a un grupo de diagnóstico y los asignados a otro.

Sea que el programa siga el enfoque clínico más popular o el enfoque estadístico-actuarial, la dificultad de las reglas o los algoritmos usados para generar informes CBTI fluctúa desde (1) un procedimiento simple en el cual una calificación o un rango de calificación dados se conectan a un conjunto de párrafos interpretativos breves, hasta (2) un complejo conjunto de reglas de decisión si-entonces donde un patrón particular de subcalificaciones conduce a una afirmación interpretativa determinada. Sin embargo, incluso las interpretaciones más complejas no son, por lo regular, tan individualizadas como las elaboradas de manera impresionista por un clínico o psicólogo orientador. Los algoritmos y árboles de decisión seguidos en varios programas CBTI que resultan en los mismos datos producen el mismo conjunto de afirmaciones verbales. El lector de un informe semejante puede quedar impresionado por su aura científica, pero también puede encontrar que las afirmaciones de interpretación son demasiado largas y repetitivas. Murphy y Davidshofer (1994) sugieren que entre más diga un informe CBTI acerca de una persona, menos probable es que sea verdad.

La mayoría de los programas CBTI están diseñados para tomar en consideración la edad, el género y otra información demográfica acerca del examinado, pero ningún programa considera todos sus atributos personales. En consecuencia, los examinadores psicológicos complementan, por lo regular, la información basada en la computadora con afirmaciones interpretativas adicionales producto de sus propias observaciones y experiencias. Los informes CBTI no son un sustituto adecuado del juicio clínico, y los clínicos capacitados deberían revisarlos y tal vez afinarlos con información obtenida de otras fuentes. Sin embargo, adornar o alterar de cualquier manera una interpretación basada en la computadora no debe hacerse de manera rutinaria, sino sólo por buenas y apremiantes razones.

OTROS INVENTARIOS DE PERSONALIDAD ADECUADOS AL CRITERIO

Al igual que la escala Stanford-Binet en la evaluación de la inteligencia, el MMPI ha sido un instrumento fuente para otros inventarios de personalidad. Uno de estos sucesores fue el Inventario de Asesoramiento de Minnesota (MCI); otro fue el Inventario Psicológico de California (CPP).

Inventario Psicológico de California

De los muchos inventarios con validez empírica similares al MMPI para individuos normales, el más popular y extensamente investigado es el Inventario Psicológico de California (CPI). Diseñado por Harrison Gough, la mitad de las 480 afirmaciones de verdadero-falso en la versión original de este inventario de características de personalidad de adolescentes y adultos fueron tomadas del MMPI y la mitad restante eran nuevas. A diferencia de las escalas clínicas del MMPI, las cuales se relacionan sobre todo con el desajuste y los trastornos psiquiátricos, las escalas del CPI evalúan aspectos más positivos y normales de la personalidad.

Escalas del CPI. El CPI original se calificó para las escalas marcadas con asterisco (*) que se presentan en la tabla 17.3, tres de las cuales, Bienestar, Buena impresión y Comunalidad, son escalas de validez. Las primeras dos escalas de validez fueron elaboradas a partir de reactivos que la gente normal tiende a responder de cierta manera cuando se le pide que simule para verse mal

TABLA 17.3 Escalas populares y especiales del Inventario Psicológico de California, tercera edición

ESCALAS POPULARES		ESCALAS E ÍNDICES ESPECIALES
*Dominio	*Comunalidad	Potencial para la administración
*Capacidad para el estatus	*Tolerancia	Orientación al trabajo
*Sociabilidad	*Bienestar	Índice de potencial para el liderazgo
*Presencia social	*Logro por medio de la conformidad	Índice de madurez social
*Aceptación de sí mismo	*Logro por medio de la independencia	Índice de potencial creativo
Independencia	*Eficiencia intelectual	
Empatía	*Atención psicológica	
*Responsabilidad	*Flexibilidad	
*Socialización	*Feminidad-Masculinidad	
*Autocontrol		
*Buena impresión		

*Escalas en la versión original del Inventario Psicológico de California.

(Bienestar) o para verse bien (Buena impresión), mientras que la calificación de Comunalidad es sólo un recuento de las respuestas muy populares. 11 de las 15 escalas restantes, como las del MMPI, fueron seleccionadas comparando las respuestas de diferentes grupos de personas; en las otras cuatro escalas (Presencia Social, Aceptación de sí mismo, Autocontrol y Flexibilidad) se validó el contenido.

CPI™ tercera edición. La primera revisión del CPI incluía 462 reactivos que se retuvieron o replantearon de la versión original del CPI con 480 reactivos; una tercera edición subsecuente contenía 434 de esos reactivos (Gough y Bradley, 1996). La tercera edición revisada del CPI™, se califica en 20 escalas populares que constan de las 18 escalas originales del CPI más Empatía e Independencia (vea la tabla 17.3). También se califica en tres escalas de vectores y 13 escalas de propósitos especiales, cinco de los cuales se presentan en la tabla 17.3. Las tres escalas de vectores representan un modelo teórico que contiene tres temas principales: papel, carácter y competencia. El primero de los tres vectores (v. 1) mide el tema del papel, u orientación interpersonal –(interior contra exterior) la presentación interpersonal del yo inherente en las escalas de capacidad para Estatus, Dominio, Aceptación de sí mismo, Sociabilidad y Presencia social–. El segundo vector (v. 2) mide el tema del carácter (favorece la norma contra cuestiona la norma), el cual involucra valores interpersonales del tipo evaluado por las escalas de Responsabilidad, Socialización y Autocontrol. El tercer vector (v. 3) mide el tema de la competencia o realización; éste constituye una combinación de calificaciones en las escalas de Logro por medio de la conformidad, Logro por medio de la independencia, Eficiencia intelectual, Bienestar y Tolerancia. Las calificaciones de los tres vectores no están correlacionadas, pero tienen una relación significativa con las calificaciones dadas en las escalas de concepto popular.

Las calificaciones en v. 1 y v. 2 se clasificaron por separado para producir una tipología cuádruple: alfa, beta, gama y delta. Los *alfa* están orientados hacia el exterior y favorecen las normas; los *beta* están orientados hacia el interior y favorecen las normas; los *gama* están orientados hacia el exterior y dudan de las normas; los *delta* están orientados hacia el interior y dudan de las normas. Los alfa son descritos como “ambiciosos, asertivos, emprendedores, sociables y

buscadores del yo"; los beta son "cautelosos, conservadores, convencionales, moderados y modestos"; los gama son "audaces, listos, voluntariosos, progresistas y rebeldes"; los delta son "preocupados, tranquilos, reservados, sensibles e inquietos".

La tercera escala estructural (v. 3) se divide en siete niveles de competencia. El nivel 1 se describe como integración pobre con poca o ninguna realización del potencial positivo del tipo; el nivel 4 es descrito como integración promedio del ego con realización moderada del potencial positivo del tipo; el nivel 7 es descrito como "integración superior del ego con buena realización del potencial positivo del tipo" (siendo el tipo alfa, beta, gama o delta). Las descripciones de los niveles 2 y 3 se encuentran entre las de los niveles 1 y 4, y las descripciones de los niveles 5 y 6 se encuentran entre las de los niveles 4 y 7. La combinación de las calificaciones en v. 1, v. 2 y v. 3 da por resultado un total de 4 tipos $\times$ 7 niveles = 28 diferentes configuraciones de personalidad.

El editor Consulting Psychologists Press ha enfatizado los usos del CPI en la identificación y el desarrollo de empleados y líderes exitosos, la creación de organizaciones eficientes y productivas, y la promoción de equipos de trabajo. Al parecer, la tipología cuádruple descrita líneas arriba es un rasgo atractivo del inventario entre los usuarios en la industria y otras organizaciones.

Inventario de Personalidad para Niños

Debido a que la mayoría de los niños tiene una baja comprensión de lectura, los inventarios de autorreporte son menos confiables y menos válidos cuando se usan con ellos. Aunque existen numerosos inventarios en donde los niños se califican a sí mismos, los instrumentos en que los adultos califican a los niños a menudo son más válidos. Un ejemplo es el Inventario de Personalidad para Niños, segunda edición (PIC-2) (por D. Lachar y C. P. Gruber; Western Psychological Services; vea Lachar, 1999). El PIC-2, que consta de 275 reactivos de verdadero-falso concernientes a la conducta del niño, puede ser completado en 40 minutos por uno de los padres, a menudo lo hace la madre u otra persona que se encargue del cuidado del niño. Las respuestas son calificadas en tres escalas de validez de respuesta (Inconsistencia, Simulación, Defensividad) y nueve escalas de ajuste (Deterioro cognoscitivo, Delincuencia, Disfunción familiar, Impulsividad y tendencia a la distracción, Incomodidad psicológica, Distorsión de la realidad, Déficit en las habilidades sociales, Aislamiento social, Preocupación somática). Cada una de las escalas de ajuste se califica en dos o tres subescalas.

Una versión más breve del PIC-2, el Resumen conductual, puede utilizarse para la detección, investigación o supervisión de cambios en el comportamiento. Consta de los primeros 96 reactivos y se califica en ocho escalas de ajuste. Otros dos instrumentos en la familia PIC-2 son el Inventario de Personalidad para los Jóvenes (PIY), un inventario de autorreporte para niños (del grado 4 al 12), y el Estudio de la Conducta del Estudiante (SBS), una escala para calificar al niño que es completada por los maestros.

El PIC-2 fue estandarizado en 2,306 padres de niños en los grados K a 12 en 23 escuelas urbanas, rurales y suburbanas en 12 estados. Se obtuvieron datos adicionales en una muestra de 1,551 padres cuyos hijos habían sido canalizados para recibir intervención educativa o clínica.

Inventario Multiaxial Clínico de Millon

El Inventario Multiaxial Clínico de Millon-III (MCMI-III) fue diseñado para evaluar los trastornos de personalidad y síndromes clínicos relacionados con el DSM-IV y fue coordinado con la teoría de la personalidad de Theodore Millon (Millon, Millon y Davis, 1994). Al revisar el MCMI-II para producir el MCMI-III, 95 de los 175 reactivos se volvieron a redactar o se reemplazaron para obtener una alineación más cercana con el DSM-IV. De las 24 escalas de diagnóstico

del MCMI-III, 14 son escalas de patrón de personalidad que se coordinan con los trastornos del Eje II del DSM-IV, y 10 son escalas de síndromes clínicos asociados con trastornos en el Eje I del DSM-IV. Además, hay tres índices de modificación y un índice de validez para detectar las respuestas descuidadas, confusas o aleatorias (vea la tabla 17.4).

El MCMI-III puede aplicarse a adultos (de 18 años en adelante) en aproximadamente 25 minutos en situaciones clínicas donde se evalúa a la gente para detectar dificultades emocionales, conductuales o personales. Las calificaciones crudas de las diversas escalas son ponderadas y convertidas a *calificaciones de tasa base*, las cuales toman en consideración la incidencia de una característica o un trastorno particular en la población general. Al determinar la ocurrencia de un trastorno o rasgo particular de personalidad en una población específica, las calificaciones crudas pueden transformarse para maximizar la razón del número de clasificaciones correctas (positivos válidos) con el número de clasificaciones incorrectas (falsos positivos).

Además de las calificaciones en el MCMI-III, NCS Assessments proporciona un informe narrativo orientado al tratamiento que incluye afirmaciones sobre la validez de las respuestas, interpretaciones para los trastornos del Eje I y el Eje II del DSM-IV, comentarios acerca de las respuestas dignas de atención dadas por los examinados, diagnósticos multiaxiales paralelos al DSM-IV e implicaciones terapéuticas. Una sección sumaria del informe reseña la gravedad de los síntomas y proporciona una breve descripción de indicaciones en los Ejes I y II del DSM-IV y consideraciones relacionadas de tratamiento.

El manual del MCMI-III describe datos de estandarización de varias muestras (Millon, Millon y Davis, 1994). La muestra normativa constó de 1,000 hombres y mujeres que representaban una amplia variedad de diagnósticos psiquiátricos. Las confiabilidades promedio test-retest y por consistencia interna de las escalas del MCMI-III son muy buenas para un inventario de personalidad. Los coeficientes alfa van de .66 a .89 para las escalas de personalidad y de .71 a .90 para las de síndrome clínico; las confiabilidades test-retest van de .84 a .96 para las escalas de síndrome clínico.

TABLA 17.4 Escalas del MCMI-III

ÍNDICES MODIFICADORES	
X	Revelación
Y	Deseabilidad
Z	Degradación
ESCALAS CLÍNICAS DEL PATRÓN DE PERSONALIDAD	
1	Esquizoide
2	Evasivo
3	Dependiente
4	Histriónico
5	Narcisista
6A	Antisocial
6B	Agresivo (sádico)
7	Compulsivo
8A	Pasivo-Agresivo (negativista)
8B	Autoderrotado (masoquista)
ESCALAS DE PATOLOGÍA GRAVE DE LA PERSONALIDAD	
S	Esquizotípica
C	Limítrofe
P	Paranoide
ESCALAS DE SÍNDROMES CLÍNICOS	
A	Trastorno de ansiedad
H	Trastorno somatomorfo
N	Bipolar: trastorno maníaco
D	Trastorno distímico
B	Dependencia del alcohol
T	Dependencia de drogas
ESCALAS DE SÍNDROMES CLÍNICOS GRAVES	
SS	Trastorno del pensamiento
CC	Depresión mayor
PP	Trastorno delirante

Fuente:

Inventario Básico de la Personalidad

El Formato de Investigación de Personalidad, descrita antes en este capítulo, y su sucesor, el Inventario de Personalidad de Jackson, fueron diseñados para evaluar personalidades normales. La orientación de otro instrumento de evaluación de la personalidad diseñado por D. N. Jackson y sus colaboradores, el Inventario Básico de la Personalidad (BPI) (por D. N. Jackson; Sigma Assessment Systems), es algo diferente. Se pretendía que el BPI se utilizara con poblaciones clínicas y normales para identificar fuentes de desajuste y fortalezas personales en jóvenes y adultos. Consta de 240 reactivos de verdadero-falso escritos a nivel de quinto grado y puede responderse en alrededor de 35 minutos. Las respuestas se califican en 12 escalas básicas: Alienación, Ansiedad, Negación, Depresión, Desviación, Hipocondriasis, Expresión de impulsos, Problemas interpersonales, Ideas persecutorias, Automenosprecio, Introversión social y Trastorno del pensamiento.

La orientación del BPI hacia el desajuste lo ha vuelto más aplicable en la práctica psicológica, psiquiátrica y de consejería en instalaciones correccionales para jóvenes y adultos y en otros contextos organizacionales interesados en los trastornos de comportamiento o conductuales. El manual proporciona perfiles separados para grupos de pacientes psiquiátricos con síntomas que van desde anorexia hasta conducta suicida y alucinaciones (Jackson *et al*., 1989). Las normas usadas para los perfiles de adultos descritos en el manual se obtuvieron a partir de encuestas por correo o entrevistas de 709 hombres y 710 mujeres seleccionados al azar de los directorios telefónicos y los registros de votantes en Estados Unidos y Canadá. Las confiabilidades por consistencia interna y test-retest de las escalas del BPI van de moderadas a altas en las muestras clínicas y no clínicas (Holden, Fekken, Reddon, Helmes y Jackson, 1988; Jackson *et al*., 1989). La información sobre la validez del BPI no es tan adecuada como podríamos desear, pero el inventario parece ser prometedor para el uso clínico en los campos de la psicología de la salud y la delincuencia juvenil en particular.

Inventario de Evaluación de la Personalidad

El Inventario de Evaluación de la Personalidad (PAI) (por L. C. Morey; Psychological Assessment Resources; vea Morey, 1999) es similar al BPI en que confía en una estrategia combinada racional-empírica para el desarrollo de instrumentos. Este inventario de autorreporte consta de 344 reactivos de cuatro puntos (F = Falso, Nada cierto; LC = Ligeramente Cierto; PC = Principalmente Cierto; MC = Muy Cierto) escritos a nivel de cuarto grado. Como una alternativa multidimensional al MMPI, el PAI fue diseñado para proporcionar información relevante para los diagnósticos clínicos, planes de tratamiento y detección de psicopatología en adultos de 18 años en adelante. Puede calificarse en cuatro escalas de validez (Inconsistencia, Infrecuencia, Impresión negativa, Impresión positiva), 11 escalas clínicas (Quejas somáticas, Ansiedad, Trastornos relacionados con la ansiedad, Depresión, Manía, Paranoia, Esquizofrenia, Rasgos limítrofes, Rasgos antisociales, Problemas con el alcohol, Problemas con las drogas), cinco escalas de tratamiento (Agresión, Ideación suicida, Estrés, Falta de apoyo, Rechazo al tratamiento), y dos escalas interpersonales (Dominio, Calidez). Diez escalas se subdividen en 31 escalas conceptualmente distintas.

Las normas estadounidenses para el PAI se basan en una muestra normativa de adultos estadounidenses de 18 años en adelante, estratificados por género, raza y edad de acuerdo con las proyecciones del censo estadounidense de 1995. También se dispone de normas en muestras

grandes de pacientes clínicos adultos y estudiantes universitarios, y en muestras de otros países y culturas.

Las revisiones del PAI han sido, por lo general, positivas (Boyle, 1995; Kavan, 1995; White, 1996), y se ha conducido una cantidad considerable de investigación con dicho instrumento en diversos grupos, particularmente en contextos clínicos y forenses (por ejemplo, Edens, Hart, Johnson, Johnson y Olver, 2000; Hays, 1997; Rogers, Ustad y Salekin, 1998; Wang *et al*., 1997). El número de citas de investigación del PAI en PsycINFO desde 1995 ha sido más alto que para cualquier otro inventario de personalidad excepto el MMPI-2. De acuerdo con el resumen de Piotrowski (2000) de los hallazgos de la investigación, el PAI se encuentra entre las pruebas objetivas de personalidad usadas con mayor frecuencia en la práctica y el entrenamiento clínico.

RESUMEN

Desde que la Hoja de Datos Personales de Woodworth hizo su aparición durante la Primera Guerra Mundial, se han elaborado numerosos inventarios de una sola calificación y de calificación múltiple. Esos inventarios miden la posición de la persona en ciertas variables de ajuste, temperamento, rasgo o psiquiátricas. Los reactivos y las escalas en algunos inventarios de personalidad se basan en un marco de referencia racional o teórico. La selección de reactivos y la calificación de otros inventarios se determinan a partir de los resultados del análisis factorial o de estudios empíricos de la capacidad de los reactivos para diferenciar entre varios grupos criterio.

Ejemplos de inventarios basados en una teoría de la personalidad son la Escala de Preferencias Personales de Edwards, el Indicador de Tipos Psicológicos de Myers-Briggs y el Formato de Investigación de la Personalidad. Entre los inventarios basados en los resultados del análisis factorial se encuentran el Estudio de Temperamento de Guilford-Zimmerman, el Cuestionario de 16 Factores de Personalidad, el Inventario de Personalidad para Adultos, el Cuestionario de Personalidad de Eysenck, el Inventario NEO de Personalidad y el Inventario NEO de Cinco Factores.

El inventario de personalidad más famoso, y sobre el cual se ha conducido la mayor cantidad de investigación, es el Inventario Multifásico de Personalidad de Minnesota (MMPI). El MMPI, un inventario con criterio codificado, fue diseñado para diferenciar entre varios grupos de diagnóstico en nueve escalas clínicas analizando las diferencias en las respuestas de gente normal y de pacientes con diagnósticos psiquiátricos específicos. El MMPI puede ser calificado también en muchas otras escalas, incluyendo las escalas de validación (*?*, *L*, *F*, *K*). Las escalas de validación en el MMPI y en otros inventarios de personalidad se califican para determinar si los reactivos han sido respondidos de manera apropiada y para ajustar calificaciones en las calificaciones de contenido por simulación y grupos de respuesta. En 1986 se publicó una versión revisada y reestandarizada del MMPI, el MMPI-2.

La interpretación de las pruebas basada en la computadora, la cual empezó con el trabajo en la Clínica Mayo a principios de la década de 1960 con la calificación interpretativa del MMPI, se ha extendido para incluir la calificación e interpretación de docenas de instrumentos cognoscitivos y afectivos por cientos de organizaciones comerciales.

El procedimiento de codificación de criterio con el cual se elaboró el MMPI también fue empleado al preparar el Inventario Psicológico de California, el Inventario de Personalidad para Niños y otros instrumentos relacionados. Otros inventarios de personalidad notables diseñados sobre la base de codificación de criterio, así como por la teoría y procedimientos psicométricos complejos, incluyen al Inventario Multiaxial Clínico de Millon-III, el Inventario Básico de Personalidad y el Inventario de Evaluación de la Personalidad.

PREGUNTAS Y ACTIVIDADES

1. Escriba cinco reactivos de verdadero-falso, concernientes a características de personalidad, que usted piense serían respondidos como "verdadero" más a menudo por simuladores que por no simuladores. Luego escriba cinco reactivos que usted crea serían respondidos como "falso" más a menudo por simuladores que por no simuladores.

2. Arregle las cosas para presentar el Indicador de Tipos de Myers-Briggs, el Inventario Psicológico de California u otro inventario de personalidad y haga que sus puntuaciones sean interpretadas por una persona calificada. ¿Sus calificaciones son consistentes con su propia evaluación de su personalidad? ¿Qué críticas del inventario puede ofrecer?

3. Mencione y describa tres inventarios de personalidad diseñados de acuerdo con cada una de las siguientes estrategias: validación de contenido, análisis factorial, codificación de criterio.

4. Elabore un inventario de autoconcepto de diez reactivos usando un formato de verdadero-falso. En cinco de las afirmaciones la respuesta codificada debe ser "verdadero" y en las otras cinco afirmaciones debe ser "falso". Aplique su inventario de autoconcepto a varios estudiantes y calcule sus calificaciones totales (de 0 a 10) de acuerdo con el número de respuestas dadas en la dirección codificada. ¿En general, las respuestas fueron bajas o altas? ¿Qué tan variables fueron? ¿Qué evidencia existe con respecto a la confiabilidad y validez de su inventario de autoconcepto?

5. Complete la Escala de calificación de personalidad de cinco variables en el formato 16.3 (página 375) y califique sus respuestas de acuerdo con el siguiente procedimiento.

   Las fórmulas de calificación para las cinco variables son:

   Agradabilidad = 5 + reactivo 12 + reactivo 15 – reactivo 8
   Escrupulosidad = 13 – reactivo 3 + reactivo 5 – reactivo 13
   Extroversión = 13 – reactivo 1 – reactivo 7 + reactivo 11
   Neuroticismo = 13 + reactivo 2 – reactivo 9 – reactivo 14
   Apertura = reactivo 4 + reactivo 6 + reactivo 10 – 3

   Ordene sus calificaciones en las cinco escalas de este inventario. ¿Son sus calificaciones más alta y más baja congruentes con la evaluación subjetiva de sus características de personalidad?

6. Haga copias múltiples del siguiente inventario de personalidad y aplíquelo a varias personas.

   Instrucciones: Por cada afirmación encierre en un círculo el número que indique qué tan cierta es en relación con usted.

¿QUÉ TAN CIERTO ES ESTO DE USTED?

		DIFÍCILMENTE				MUCHO
1.	Hago amigos con facilidad.	1	2	3	4	5
2.	Tiendo a ser tímido.	1	2	3	4	5
3.	Me gusta estar con otros.	1	2	3	4	5
4.	Me gusta ser independiente de la gente.	1	2	3	4	5
5.	Por lo general prefiero hacer las cosas solo.	1	2	3	4	5
6.	Siempre estoy en movimiento.	1	2	3	4	5
7.	Me gusta salir y correr tan pronto como me despierto en la mañana.	1	2	3	4	5
8.	Me gusta mantenerme ocupado todo el tiempo.	1	2	3	4	5

9.	Tengo mucha energía.	1	2	3	4	5
10.	Prefiero los pasatiempos tranquilos e inactivos a los más activos.	1	2	3	4	5
11.	Tiendo a llorar con facilidad.	1	2	3	4	5
12.	Me asusto con facilidad.	1	2	3	4	5
13.	Tiendo a ser algo emocional.	1	2	3	4	5
14.	Me enfado con facilidad.	1	2	3	4	5
15.	Tiendo a irritarme con facilidad.	1	2	3	4	5

(Tomado de *Individual and Group Differences*, por L. Willerman, Nueva York: Harper's College Press, 1975. Reproducido con autorización.)

Calificación: Para todos los reactivos, excepto los números 2, 4, 5 y 10, la calificación es simplemente el número encerrado en un círculo; para los reactivos 2, 4, 5 y 10, invierta los números antes de calificar (1 = 5, 2 = 4, 3 = 3, 4 = 2, 5 = 1). Sume las calificaciones en los reactivos 1 a 5 para obtener un índice de sociabilidad; sume las calificaciones en los reactivos 6 a 10 para obtener un índice de nivel de actividad; sume las calificaciones en los reactivos 11 a 15 para obtener un índice de emocionalidad. Interprete sus calificaciones y las de las otras personas a quienes aplicó el inventario con respecto a las siguientes normas. Estas normas constituyen los rangos dentro de los cuales 68% de las muestras de estudiantes de ambos sexos de una gran universidad calificaron en las tres escalas.

RANGO ÍNDICE	RANGO PROMEDIO PARA HOMBRES	PROMEDIO PARA MUJERES
Sociabilidad	13–19	15–20
Nivel de actividad	13–19	13–20
Emocionalidad	11–18	9–16

Las calificaciones fuera de esos rangos pueden interpretarse como alta o baja en la característica particular. ¿Por qué es importante que sea cauteloso en sus interpretaciones de las calificaciones en este inventario?

7. ¿Cuáles son las diferencias entre inventarios de personalidad, escalas de calificación y listas de verificación en términos de sus usos, formato, calificación y cualidades psicométricas?

8. En los años recientes ha existido una gran discusión en la literatura psicológica concerniente a la variable de optimismo-pesimismo. Trate de elaborar un inventario de optimismo-pesimismo que conste de diez reactivos. Cinco reactivos deberán estar redactados en una dirección positiva (optimista) y los otros cinco en una dirección negativa (pesimista). Usando cualquier formato de respuesta que desee y las instrucciones apropiadas, denomine a su inventario: "Cuestionario de actitudes". Mecanografíe su inventario en un solo lado de una hoja de papel, haga varias copias y aplíquelas al mismo número de personas. Obtenga comentarios sobre su inventario de los examinados y de otras personas, y haga los cambios necesarios en la redacción o el formato.

CAPÍTULO DIECIOCHO

TÉCNICAS PROYECTIVAS

Lawrence Frank (1939) acuñó el término *técnica proyectiva* para referirse a los procedimientos de evaluación psicológica en los cuales las personas "proyectan" sus necesidades y sentimientos internos en estímulos ambiguos. Los estímulos son materiales y/o tareas relativamente no estructurados en los que se pide a la persona describir, contar una historia al respecto, completar o responder de alguna otra forma. En contraste con instrumentos más directos como los inventarios de personalidad y las escalas de calificación, las técnicas proyectivas son, por lo general, menos obvias en su propósito y, en consecuencia, se supone que están menos sujetas a la simulación y los grupos de respuesta. Como los materiales o tareas de estímulo son de un contenido relativamente no estructurado y flexibles en términos de la respuesta provocada, se supone que la estructura impuesta por la persona que responde es un reflejo, o *proyección*, de sus percepciones individuales de las cosas. También se supone que los materiales menos estructurados tienen mayor probabilidad de revelar facetas importantes de la personalidad que los más estructurados.

Las aplicaciones de las técnicas proyectivas son similares a las de los inventarios de personalidad, pero las técnicas proyectivas se usan más con propósitos de psicodiagnóstico en clínicas de salud mental, hospitales y centros de orientación. Sin embargo, las técnicas proyectivas no proporcionan un "ábrete sésamo" al inconsciente o una radiografía de la mente. Además, la falta de estructura es un arma de dos filos que puede dar como resultado una gran cantidad de datos difíciles de interpretar.

Debido a los problemas de calificación, la mayoría de las técnicas proyectivas no logra cumplir los estándares convencionales de confiabilidad y validez. Sus coeficientes de validez son generalmente bajos, lo cual refleja factores situacionales y de subjetividad en la calificación e interpretación. En comparación con los usuarios de los inventarios de personalidad, a los usuarios de las técnicas proyectivas tiende a preocuparles menos la confiabilidad, la validez y las normas, interesándoles más la riqueza de la interpretación impresionista y el análisis clínico de las respuestas.

Como las técnicas proyectivas intentan capturar un proceso inconsciente, la interpretación de las respuestas que suscitan ha recibido una gran influencia de la teoría psicoanalítica. Esto es cierto, sobre todo, en las Ilustraciones de Blacky, una técnica de narración e imágenes basada en la teoría freudiana de las etapas psicosexuales del desarrollo (Blum, 1949, 1950).

No es sorprendente que el mayor incremento en el uso de técnicas proyectivas tuviera lugar entre 1940 y 1960, una época en que el pensamiento psicoanalítico ejerció una influencia muy fuerte en la teoría y la investigación de la personalidad. Muchas técnicas proyectivas disponen de sistemas formales de calificación, pero los psicólogos clínicos y orientadores que interpretan los protocolos de las pruebas proyectivas tratan, por lo general, de formarse una impresión global de la personalidad del examinado buscando consistencias y rasgos sobresalientes en el patrón de respuestas. En lugar de conducir a un diagnóstico preciso, las respuestas a los

materiales proyectivos pueden sugerir hipótesis explicatorias concernientes a la psicodinámica de la personalidad y los problemas del examinado.

Se dice que la aplicación y calificación de una técnica proyectiva típica requiere mayor capacitación y sensibilidad de la necesaria para un inventario de autorreporte. Aun así los psicólogos, que se supone están bien capacitados en las técnicas proyectivas, con frecuencia discrepan en sus interpretaciones de las respuestas recibidas. Las interpretaciones descritas en las siguientes secciones, y el análisis de las respuestas a las técnicas proyectivas en general, deben verse como especulaciones o posibilidades más que como un hecho confirmado. Por ejemplo, considere las siguientes interpretaciones sugeridas para los dibujos de una figura:

> Las figuras de gran tamaño se interpretan como generosidad emocional o conducta de expresión de impulsos reprimidos (*acting-out*).
>
> Las figuras de tamaño pequeño se interpretan como restricción emocional, aislamiento o timidez.
>
> Los borrones alrededor de los glúteos y/o pestañas largas en la figura masculina se interpretan como homosexualidad latente.
>
> El detalle excesivo en las líneas se interpreta como tensión y conducta agresiva y posible pensamiento delirante.
>
> La distorsión u omisión de los rasgos faciales u otros rasgos del cuerpo se interpretan como un conflicto relacionado con la característica u órgano correspondiente.

Estas interpretaciones pueden parecer aceptables, pero se basan en gran medida en estereotipos o correlaciones ilusorias y a menudo resultan más erróneas que ciertas. Por supuesto, concluir que algo es cierto sólo porque parece razonable es un error que no se limita a quienes elaboran diagnósticos clínicos. Las interpretaciones de las respuestas a las técnicas proyectivas deben verse sólo como posibilidades o hipótesis razonables que pueden o no ser confirmadas por otras fuentes de información concernientes a la persona.

ELABORACIONES Y ASOCIACIONES DE PALABRAS

Se han desarrollado varias técnicas proyectivas para detectar motivos poco obvios, conflictos, problemas y otras características intrapersonales encubiertas. De éstas, técnicas semiestructuradas como las asociaciones de palabras y las frases incompletas son las que quizá estén más cercanas a los inventarios de autorreporte en términos de diseño, formato y objetividad.

Asociaciones de palabras

El método de asociación de palabras fue introducido por Francis Galton (1879), y Carl Jung (1910) efectuó la primera aplicación clínica para detectar conflictos neuróticos. En este procedimiento se lee en voz alta una serie de palabras a la persona a examinar, a quien se le ha indicado responder a cada término con la primera palabra que le venga a la mente. Las aplicaciones clínicas de la técnica implican intercalar dentro de un conjunto de términos neutrales palabras seleccionadas que contienen carga emocional o un significado especial para la persona. Además de asociaciones verbales significativas y demoras al responder, es posible determinar el grado en que ciertas palabras activan emoción midiendo la conductividad de la piel, la tensión muscular, la tasa de respiración, la presión sanguínea, la tasa del pulso, el temblor de la voz u otras reac-

ciones fisiológicas ante las palabras que sirven de estímulo. Las palabras también pueden ser usadas como estímulos en una prueba de polígrafo (detector de mentiras) diseñada para detectar respuestas emocionales altas a ciertos acontecimientos concernientes a un delito sobre el cual la persona tiene cierto conocimiento.

Al igual que con todas las técnicas proyectivas, las respuestas en una prueba de asociación de palabras deben interpretarse en un contexto donde se cuente con otra información sobre la persona. Un principio general que ha guiado las interpretaciones psicoanalíticas del lenguaje es que los sustantivos tienen mayor probabilidad que los verbos de ser expresiones disfrazadas de necesidades y conflictos. Esto es así porque, de acuerdo con la teoría freudiana, es más fácil alterar el objeto de un deseo (un sustantivo) que su dirección (un verbo).

Muchos psicólogos clínicos prefieren elaborar sus propias listas de palabras, pero se dispone de listas estandarizadas. Un ejemplo es la Prueba Kent-Rosanoff de Asociación Libre, una lista estandarizada de 100 palabras y las asociaciones correspondientes dadas por 1,000 adultos. La Kent-Rosanoff, publicada originalmente en 1910, es una de las pruebas psicológicas más antiguas en uso (por ejemplo, Isaacs y Chen, 1990). Otra lista estandarizada de palabras fue proporcionada por Rapaport, Gill y Schafer (1968) y se utiliza con propósitos de diagnóstico en la Clínica Menninger.

Frases incompletas

Pedir a una persona que complete enunciados inconclusos preparados de manera especial es una técnica proyectiva flexible y de fácil aplicación descrita en un principio por Payne (1928). Es posible elaborar una variedad de fragmentos de enunciados o *troncos* relacionados con posibles áreas de activación emocional y conflicto. Algunos ejemplos son los siguientes:

Mi mayor preocupación es _____.
Sólo desearía que mi madre hubiera _____.
La cosa que más me molesta es _____.
No me gusta _____.

Se supone que los deseos, anhelos, temores y actitudes de la persona se reflejan en la forma en que completa las frases.

A pesar de que son más obvias que muchas otras técnicas proyectivas, las frases incompletas se consideran una de las técnicas proyectivas más válidas para propósitos de diagnóstico e investigación (Haak, 1990; Lah, 1989). La confiabilidad y validez de las frases incompletas son más altas cuando las respuestas se califican e interpretan de manera objetiva que de manera impresionista. Como con el MMPI y otros instrumentos de criterio codificado, se han desarrollado claves empíricas para las asociaciones de palabras y las frases incompletas.

Las pruebas de frases incompletas pueden ser elaboradas para un caso clínico particular o una investigación de la personalidad, pero en el mercado se dispone de varios instrumentos de este tipo. Algunos ejemplos son la Serie de Frases Incompletas y la Técnica de Frases Incompletas EPS (ambas de Psychological Assessment Resources) y el Formulario Rotter de Frases Incompletas, segunda edición (por J. B. Rotter, Psychological Corporation).

Formulario de Frases Incompletas de Rotter, segunda edición (RISB). Cada una de las tres formas del RISB (bachillerato, universidad, adulto) consta de 40 fragmentos de enunciados escritos sobre todo en primera persona y requiere de 20 a 40 minutos para completarse. Las respuestas se califican para conflicto (C) o respuestas no saludables (por ejemplo, “Odio.... casi a

todos"); respuesta positiva (P) (por ejemplo, "Lo mejor... está por venir"), y respuesta neutral (N) (por ejemplo, "La mayoría de las chicas... son mujeres"). No responder o dar una respuesta demasiado corta para ser significativa se cuenta como omisión. Los pesos de calificación para las respuestas C son C1 = 4, C2 = 5 y C3 = 6, del más bajo al más alto grado de conflicto expresado. Los pesos de calificación para las respuestas P son P1 = 2, P2 = 1 y P3 = 0, de la respuesta menos a la más positiva. Las respuestas N no reciben pesos numéricos. Después de calificar las respuestas a todos los fragmentos de los enunciados, se obtiene una calificación de ajuste global sumando las calificaciones ponderadas en las categorías de conflicto y positiva. La calificación de ajuste global va de 0 a 240, y las calificaciones más altas se asocian con mayor desajuste. La segunda edición (1992) del manual del RISB contiene algunos ejemplos de casos para demostrar la calificación de este instrumento.

La investigación ha demostrado que el RISB permite clasificar correctamente a la mayoría de los individuos en categorías de ajuste y desajuste y, por ende, puede ser usado para detectar desajuste global. Las últimas normas (1992) proporcionan calificaciones límite basadas en muestras ajustadas y no ajustadas. El manual de 1992 también contiene una revisión actualizada de investigaciones que comprueban la confiabilidad, validez y utilidad clínica del RISB (Rotter, Lah y Rafferty, 1992).

Estudio de Frustración Ilustrado de Rosenzweig

Otro instrumento proyectivo en el cual los examinados elaboran respuestas verbales a estímulos parcialmente verbales es el Estudio de Frustración Ilustrado de Rosenzweig (por S. Rosenzweig, Psychological Assessment Resources). Cada una de las tres formas de este instrumento

FIGURA 18.1 Reactivo del Estudio de Frustración Ilustrado de Rosenzweig.

(Derechos reservados 1964 por Saul Rosenzweig. Reproducido con autorización.)

(niño, adolescente, adulto) consta de 24 caricaturas que representan a individuos en situaciones frustrantes (vea la figura 18.1). Se pide al examinado que indique, escribiendo en el recuadro en blanco que aparece sobre la cabeza de la persona frustrada, una respuesta verbal que pudiera haber sido dada por esta persona anónima (figura 18.1). Las respuestas son calificadas de acuerdo con la dirección y el tipo de agresión expresada. En la dirección de la agresión se incluyen la extraagresiva (hacia el exterior o hacia el ambiente), la intraagresiva (hacia el interior, hacia uno mismo) y la imaagresiva (evitación o no expresión de la agresión). En el tipo de agresión se incluyen el dominio del obstáculo u O-D (el objeto frustrante destaca), defensa del ego o E-D (el ego del examinado predomina para defenderse) y la persistencia de la necesidad o N-P (la meta es perseguida a pesar de la frustración). Las calificaciones se interpretan en términos de la teoría de la frustración y de las normas disponibles del instrumento. El Estudio de Frustración Ilustrado de Rosenzweig ha sido utilizado por todo el mundo en un gran número de investigaciones interesadas en la naturaleza de la frustración y su relación con otras variables (vea Rosenzweig, 1978).

Dibujos proyectivos

Los procedimientos que requieren respuestas orales o escritas a palabras y enunciados son sólo una de las muchas tareas de elaboración caracterizadas como técnicas proyectivas. Otros materiales no verbales que han sido empleados son pinturas de arcilla, materiales de construcción y trozos de colores. El análisis de la escritura, aunque no ha recibido gran aceptación entre los psicólogos, también tiene seguidores (Holt, 1974). Instrumentos aún más populares han sido la Prueba del Dibujo de una Persona (Machover, 1971) y la Técnica de Casa-Árbol-Persona (Buck, 1992), los cuales requieren que los examinados realicen dibujos de personas o de objetos.

Prueba del Dibujo de una Persona (DPT). En esta prueba el sujeto dibuja personas de su mismo sexo y del sexo opuesto. Los dibujos se interpretan en términos de la ubicación de las diversas características del dibujo (sexo, clase, posición, ropas, etc.). Los seguidores de la técnica sostienen que la gente tiene tendencia a proyectar impulsos que le resultan aceptables en la figura del mismo sexo y los que le resultan inaceptables en la figura del sexo opuesto. Se considera que los aspectos particulares de los dibujos son indicadores de ciertas características de personalidad o condiciones psicopatológicas. Pestañas largas en las personas dibujadas indican histeria; muchos detalles en las ropas sugieren neurosis; los dibujos grandes señalan la expresión de los impulsos, y el sombreado oscuro y pesado sugiere fuertes impulsos agresivos. Los dibujos pequeños, pocos rasgos faciales o expresión facial abatida señalan depresión; pocos detalles de la periferia del cuerpo indican tendencias suicidas y pocos rasgos físicos sugieren psicosis o daño cerebral orgánico (Kaplan y Sadock, 1995). Machover (1971) sostenía que una cabeza desproporcionadamente grande o pequeña –el centro del poder intelectual, control de los impulsos del cuerpo y balance social– es indicativa de dificultades funcionales en esas áreas. Aunque se reporta que muchos de esos signos y generalizaciones interpretativos se basan en la experiencia clínica y pueden tener sentido psicoanalítico e incluso sentido común, no han pasado por un escrutinio estricto. Se ha encontrado alguna evidencia sobre una relación entre la cualidad juzgada del dibujo y el ajuste psicológico general (Lewinsohn, 1965; Roback, 1968), pero la investigación no ha apoyado la mayoría de las interpretaciones psicodinámicas de Machover y otros.

Naglieri, McNeish y Bardos (1991) desarrollaron Dibujar una Persona (DAP): Procedimiento de Detección de Perturbación Emocional para ser usado como prueba de detección por niños sospechosos de trastorno conductual y perturbación emocional. Este enfoque a la calificación del DAP se ha cumplido con cierto éxito en el diagnóstico de niños con problemas (Naglie-

ri y Pfeiffer, 1992). Se han publicado otras variaciones de la prueba DAP, como lograr que la persona dibuje varios objetos, un grupo de personas, o una historia. Dos de tales variaciones son la Técnica de Casa-Árbol-Persona y el Sistema de Dibujo Cinético para la Familia y la Escuela. El primer instrumento (por J. N. Buck y W. L. Warren; Western Psychological Services) requiere que la persona dibuje a pulso una casa, un árbol y una persona. Los dibujos pueden calificarse de manera cuantitativa en una serie de variables, pero por lo general se interpretan de manera impresionista y holística.

Sistema de Dibujo Cinético para la Familia y la Escuela. En esta técnica proyectiva (por H. M. Knoff y H. T. Prout; Western Psychological Services) se pide primero al niño o adolescente examinado que dibuje a su familia haciendo algo. Después que el dibujo ha sido terminado, se pide al examinado que identifique a cada miembro de la familia en el dibujo, lo que hacen en la ilustración y por qué, y que hablen entre sí acerca de sus relaciones. En la parte de la escuela se sigue un procedimiento similar. Al interpretar los dibujos, el examinador intenta aclarar sus significados y determinar qué procesos ocultos pueden haber influido en su elaboración. Dos de los rasgos de los dibujos de mayor importancia psicológica son el grado de interacción entre las figuras y la medida en que éstas interactúan.

PRUEBAS DE MANCHAS DE TINTA

Las manchas de tinta son uno de los tipos menos estructurados de material de pruebas proyectivas, y por ello se supone que dan rienda suelta a la expresión de los motivos y deseos inconscientes. El psiquiatra suizo Hermann Rorschach no fue el primero en utilizar las manchas de tinta para estudiar la personalidad, pero proporcionó el primer conjunto de manchas de amplia aceptación y una aproximación estándar a la aplicación e interpretación de las respuestas.

Técnica de Psicodiagnóstico de Rorschach

Los materiales de estímulo para administrar la Técnica de Psicodiagnóstico de Rorschach (de Hogrefe y Huber), publicados por primera vez en 1921, son diez tarjetas de $5^1/_2 \times 9^1/_2$ pulgadas. Cada tarjeta contiene una mancha de tinta simétrica de manera bilateral, en blanco y negro (cinco tarjetas), rojo y gris (dos tarjetas) o de muchos colores (tres tarjetas) contra un fondo blanco similar a la que se muestra en la figura 18.2. Las tarjetas se presentan de manera individual y son vistas a una distancia no mayor de la extensión del brazo; se permite voltearlas. Se pide a los examinados que informen lo que ven en la mancha o lo que puede representar. Una joven que cursaba el último año en la universidad dio la siguiente respuesta a la mancha de tinta de la figura 18.2 diez segundos después de que le fue mostrada:

> Mi primera impresión fue un gran insecto, quizá una mosca. En el fondo veo dos figuras similares a un rostro mirándose entre sí como si estuvieran hablando. También parece un esqueleto, el área de la pelvis. Veo un pequeño murciélago justo en el centro. La mitad superior parece un ratón.

Después de que se han presentado todas las tarjetas, el examinador puede empezar con la primera tarjeta y preguntar al examinado qué rasgos (forma, color, sombreado, etc.) determina-

FIGURA 18.2 Mancha de tinta similar a las de la Técnica de Psicodiagnóstico de Rorschach.

ron sus respuestas. Después de esta fase de *indagación*, puede haber otro periodo de *prueba de los límites* para descubrir si el examinado puede ver ciertas cosas en las tarjetas.

Se han propuesto varios procedimientos de calificación para la prueba de Rorschach, siendo uno de los más influyentes el amplio sistema elaborado por John Exner (1991, 1993).[1] Cada respuesta dada a una mancha puede ser calificada en varias categorías:

Ubicación: Secciones de la mancha que determinaron la percepción: la mancha entera (*W*), un detalle común (*D*), un detalle poco común (*Dd*), o, si se utilizó el espacio en blanco de la tarjeta, *WS*, *DS* o DdS.

Determinante: Rasgos de la percepción que determinaron la respuesta: forma (*F*), color (*C*), textura del sombreado (*T*), dimensión del sombreado (*V*), sombreado difuso (*Y*), color cromático (*C*), color acromático (*C'*), movimiento (*M*) o combinaciones de éstos.

Contenido: Anatomía (*An*), sangre (*Bl*), nubes (*Cl*), fuego (*Fi*), geografía (*Ge*), naturaleza (*Na*), etcétera.

Popularidad: Si la respuesta es popular (*P*) u original (*O*).

El número de respuestas en cada categoría y ciertas razones derivadas de ellas guían la interpretación del protocolo de la prueba como un todo. Por ejemplo, varias buenas respuestas de "todo" (*W*) se consideran indicativas de pensamiento integrado u organizado, mientras que las respuestas de color sugieren emocionalidad e impulsividad; se dice que un gran número de respuestas detalladas indica compulsividad; se supone que las respuestas a los espacios en blanco

[1]Se han desarrollado varios programas de cómputo para calificar e interpretar las respuestas a la prueba Rorschach. Un ejemplo es el Programa de Asistencia en la Interpretación de la prueba Rorschach™ (RIAP4™) versión 4 para Windows (por J. E. Exner, Jr. e I. B. Weiner; Psychological Assessment Resources).

señalan tendencias de oposición, y que las respuestas de movimiento revelan imaginación. Se dice que la razón del número de respuestas de movimiento humano con el número de respuestas de color (*balance de la experiencia*) está relacionada con el grado en que una persona es reflexiva más que orientada a la acción. La razón del número de respuestas de forma con el número de respuestas de color es un índice del grado en que la persona es controlada por la cognición más que por la emoción. Al evaluar el protocolo del Rorschach también es importante la precisión de las respuestas, es decir, qué tan bien se ajustan a las partes respectivas de las manchas (bien, poco e indeterminado). Las demoras al responder pueden interpretarse como ansiedad, un pequeño número de respuestas de color y movimiento como depresión, y varias respuestas de sombreado como autocontrol. Muchas respuestas originales que tienen forma deficiente y otros indicadores de pensamiento confuso sugieren un proceso psicótico.

Una de las calificaciones más confiables en la técnica de Rorschach, así como un índice aproximado de la habilidad mental, es un simple conteo del número total de respuestas a las diez manchas de tinta. Las respuestas también pueden ser interpretadas en términos del contenido, pero el proceso es muy subjetivo. Por ejemplo, los personajes irreales como los fantasmas y los payasos se interpretan como la incapacidad para identificarse con la gente real. Las máscaras se interpretan como una representación de papeles para evitar la exposición; la comida se interpreta como necesidades de dependencia o hambre emocional; la muerte como soledad y depresión, y los ojos como sensibilidad a la crítica.

Se han publicado miles de artículos sobre la técnica de Rorschach, pero no le ha ido bien en términos de su confiabilidad y validez. Considerando la cantidad de tiempo requerido para aplicar y calificar la prueba, resulta insatisfactoria cuando se juzga con criterios psicométricos convencionales.[2] Sin embargo, sigue siendo popular entre los psicólogos clínicos y los psiquiatras, y es probable que lo siga siendo hasta que se desarrolle un método que demuestre ser superior para el análisis a profundidad de la personalidad.

Técnica de Manchas de Tinta de Holtzman

La Técnica de Manchas de Tinta de Holtzman (HIT) (por W. H. Holtzman; The Psychological Corporation) representa un intento por desarrollar una prueba de manchas de tinta más objetiva y válida que la técnica de Rorschach. Las dos formas paralelas de la HIT (A y B) constan cada una de 45 manchas, y el examinado está limitado a una respuesta por mancha. Cada mancha fue seleccionada sobre la base de una elevada confiabilidad de división por mitades y por la capacidad para diferenciar entre respuestas normales y patológicas. Las manchas de la HIT son más variadas que las de Rorschach; algunas son asimétricas y otras tienen colores y diferentes texturas visuales. La HIT puede ser calificada en 22 categorías de respuesta desarrolladas por el análisis de computadora de cientos de protocolos de la prueba. Las normas por rangos percentilares para esas 22 calificaciones se basan en ocho grupos de personas, normales y patológicas, cuya edad fluctúa entre los cinco años y la adultez.

Los procedimientos mediante los que se desarrolló y estandarizó la HIT fueron más parecidos a los de un inventario de personalidad que a los de otras técnicas proyectivas, por lo que no resulta sorprendente que su confiabilidad sea mayor que la de la técnica Rorschach. Sin embargo, al igual que con ésta, todavía queda por realizar una gran cantidad de trabajo sobre la validez

[2] Weiner (1996) ofreció una evaluación más positiva de la prueba Rorschach. Concluyó que la dura crítica a este instrumento pudo haber estado justificada hasta cierto grado en los años pasados, pero "se opone a los abundantes datos contemporáneos que demuestran su solidez psicométrica y su utilidad práctica" (p. 206).

de la HIT. A pesar de sus limitaciones, la HIT es uno de los pocos instrumentos en esta categoría que se acercan a cumplir los estándares psicométricos de una buena prueba.

A finales de la década de 1980 se publicó la HIT 25, la cual requiere dos respuestas para cada una de las 25 tarjetas seleccionadas de la forma A (Holtzman, 1988). Esta variante de la HIT ha demostrado ser prometedora en el diagnóstico de la esquizofrenia ya que clasificó de manera correcta a 26 de 30 esquizofrénicos y a 28 de 30 estudiantes universitarios normales (Holtzman, 1988; vea también Swartz, 1992).

EL TAT Y VARIACIONES

Las ilustraciones y otros materiales sobre los que se pide a los individuos narrar una historia son menos estructuradas que las asociaciones de palabras y las frases incompletas, pero más estructuradas que las manchas de tinta. La mayoría de esas *pruebas de apercepción* emplea ilustraciones de personas o de animales como estímulos, pero una está compuesta por ilustraciones de manos (la Prueba de las Manos) y otra consiste en estímulos auditivos (la Prueba de Apercepción Auditiva). Casi todas las pruebas de apercepción requieren respuestas flexibles, pero al menos una, la Interpretación de Ilustraciones Iowa, tiene un formato de opción múltiple. Las instrucciones para las diversas pruebas de ilustración-historia son similares: se pide al examinado que cuente una historia acerca de cada ilustración, incluyendo lo que sucede en el momento, lo que llevó a ello y el posible resultado.

Test de Apercepción Temática

Después del Rorschach, la técnica proyectiva que le sigue en popularidad en términos de citas de investigación y uso clínico es el Test de Apercepción Temática (TAT) (por H. A. Murray; Harvard University Press). El TAT consta de 30 tarjetas con ilustraciones en blanco y negro (cuatro conjuntos traslapados de 19 tarjetas para niños, niñas, hombres y mujeres) que presentan a personas en situaciones ambiguas, más una tarjeta en blanco. La administración del TAT comienza pidiendo al examinado que cuente una historia completa acerca de cada una de las alrededor de 10 ilustraciones seleccionadas como apropiadas para su edad y sexo. Se le dice que dedique aproximadamente cinco minutos a cada historia, diciendo lo que sucede en el presente, qué pensamientos y sentimientos tiene la gente en la historia, qué acontecimientos llevaron a la situación y cómo terminará ésta. Por ejemplo, una de las ilustraciones muestra a una mujer joven en primer plano y en el fondo a una misteriosa anciana con un chal sobre la cabeza y haciendo muecas. La siguiente historia fue narrada por una joven universitaria en respuesta a esta ilustración:

> Ésta es una mujer que ha estado muy atormentada por los recuerdos de una madre hacia la que estaba resentida. Tiene sentimientos de pena por la forma en que trató a su madre; los recuerdos de su madre la atormentan. Esos sentimientos parecen aumentar conforme envejece y ve que sus propios hijos la tratan de la misma manera que ella trató a su madre. Intenta comunicar el sentimiento a sus hijos, pero no logra cambiar sus actitudes. Está viviendo el pasado en su presente, porque este sentimiento de pena y culpa es reforzado por la forma en que sus hijos la tratan.

A partir de historias como ésta se informa que se obtiene información concerniente a las necesidades, emociones, sentimientos, complejos y conflictos dominantes de la persona que las narra, así como de las presiones a las que está sometida. Como lo sugiere la historia anterior, las res-

puestas a las ilustraciones del TAT pueden ser particularmente útiles para comprender las relaciones y dificultades entre la persona y sus padres.

Al interpretar las historias del TAT se asume que las personas proyectan sus propias necesidades, deseos y conflictos en las historias y los personajes. El procedimiento tradicional de interpretación es un proceso impresionista bastante subjetivo que se centra en un análisis de las necesidades y personalidad del personaje central (el *héroe* o *heroína*), quien presumiblemente representa al examinado, y las fuerzas del entorno (*presión*) que interfieren con él. En la interpretación se toman en cuenta la frecuencia, intensidad y duración de la historia.

Los siguientes son ejemplos de las respuestas del TAT o signos que ciertos psicólogos consideran indicativos de trastornos mentales de varios tipos: lentitud o retraso para responder sugiere depresión; historias de hombres que implican comentarios negativos acerca de las mujeres o afecto por otros hombres son indicativos de homosexualidad; cautela excesiva y preocupación por los detalles son signos del trastorno obsesivo-compulsivo.

Aunque los métodos usuales de calificación e interpretación de las historias del TAT son muy impresionistas, las calificaciones determinadas por uno de los procedimientos más sistemáticos para registrar y analizar las historias son bastante confiables y pueden interpretarse en términos de normas basadas en estudios de estandarización (Bellak, 1993). Pedir a una persona que cuente historias acerca de ilustraciones también parecería tener una validez potencialmente mayor en la evaluación de la personalidad que pedir respuestas a las manchas de tinta. Sin embargo, el contenido de las historias del TAT está influido por el contexto ambiental particular en el cual se aplica la prueba, y ésta no siempre distingue entre personas normales y personas con trastornos mentales (Eron, 1950). Muchos psicólogos sostienen que los estímulos amorfos como las manchas de tinta son más efectivos que las historias sobre ilustraciones para revelar conflictos inconscientes y deseos reprimidos, pero esta afirmación nunca se ha verificado de manera adecuada. La validez de las historias sobre ilustraciones se disputa menos que la de las respuestas a las manchas de tinta, pero aún así el TAT es menos popular que la técnica de Rorschach.

Modificaciones del TAT

El TAT ha sido utilizado con una gama de grupos étnicos y de edad cronológica, y se han desarrollado varias modificaciones para no blancos, niños y adultos mayores. De acuerdo con la suposición de que los negros se identifican de manera más cercana con ilustraciones de otros negros que de blancos, 21 de las ilustraciones originales del TAT se volvieron a dibujar con personajes negros y fueron publicadas como la Modificación Thompson del TAT (Thompson, 1949). También llama la atención TEMAS (Tell-Me-a-Story), prueba diseñada de manera específica para usar con niños hispanos de la ciudad (Costantino, 1978; Costantino, Malgady y Rogler, 1988). Las 23 ilustraciones cromáticas de TEMAS representan a personajes hispanos interactuando en escenarios urbanos que implican emociones, cogniciones y actividades interpersonales negativas y positivas.

Otras dos versiones especiales del TAT son el Test de Apercepción para Personas Mayores (Bellak y Bellak, 1973) y el Test de Apercepción para Niños (Bellak y Bellak, 1949).

Test de Apercepción para Personas Mayores. Esta prueba, diseñada de manera específica para adultos mayores, consta de 16 ilustraciones acerca de las cuales se pide a los examinados relatar historias. Las ilustraciones reflejan temas de soledad, inutilidad, enfermedad, desesperanza y autoestima disminuida, además de situaciones positivas y más felices. Como en el caso de la Prueba de Apercepción Gerontológica (Wolk y Wolk, 1971), un instrumento similar, las respuestas a las ilustraciones en el Test de Apercepción para Personas Mayores refleja preocu-

paciones serias acerca de la salud, llevarse bien con otras personas y ser colocado en un asilo o una casa de retiro. Ambas pruebas han sido criticadas por normas inadecuadas y por la posible formación de estereotipos sobre la vejez.

Test de Apercepción para Niños. Con base en la suposición de que los niños pequeños (de 3 a 10 años) tienen una identificación más cercana con animales que con los seres humanos, el Test de Apercepción para Niños (CAT-A) consta de 10 ilustraciones de animales en situaciones diversas. El Test de Apercepción para Niños–Figuras Humanas (CAT-H), una extensión del CAT-A para niños mayores, consta de ilustraciones de seres humanos en situaciones paralelas a las de las ilustraciones con animales del CAT-A. Las historias tanto en el CAT-A como en el CAT-H son interpretadas desde el punto de vista de la teoría psicodinámica, de manera específica en términos de conflictos, ansiedad y culpa. Se cuenta con una lista de verificación, el Programa de Mecanismos Adaptativos de Haworth, para ayudar en la interpretación de las historias del CAT-A y el CAT-H.

OTRAS PRUEBAS DE APERCEPCIÓN

Por desgracia, la falta de representatividad y variedad de los materiales de estímulo y la carencia de rigor psicométrico en el diseño, estandarización y validación por la que se ha criticado al TAT también se aplican a las variaciones y modificaciones descritas en los párrafos anteriores. Algo más sólido desde un punto de vista psicométrico que el Test de Apercepción para Niños son el Test de Apercepción para Niños de Roberts (McArthur y Roberts, 1982) y el Test Aperceptivo de Relato de Cuentos para Niños (Schneider, 1989).

Test de Apercepción para Niños de Roberts (RATC). Esta prueba (WPS) fue diseñada para niños de 6 a 15 años, pero puede usarse también con las familias. Las 27 tarjetas de estímulo (dibujos de línea de adultos y niños en ropas modernas) en la prueba enfatizan las situaciones interpersonales cotidianas, incluyendo confrontación familiar, conflicto de los padres, afecto de los padres, observación de la desnudez y acontecimientos interpersonales en la escuela y con los compañeros, además de situaciones del tipo encontrado en el TAT y el CAT.

El RATC se aplica en dos conjuntos traslapados de 16 tarjetas cada uno, un conjunto para niños y otro para niñas. Se proporcionan directrices explícitas para calificar las historias con respecto al funcionamiento adaptado e inadaptado en las siguientes escalas: Confianza en los otros, Apoyo de los otros, Identificación del problema, Problemas no resueltos, Ansiedad, Apoyo al niño, Establecimiento de límites, Resolución, Agresión, Depresión y Rechazo. Otras tres dimensiones: Respuesta atípica, Resultado inadaptado y Referencia, sirven como indicadores críticos. Las puntuaciones crudas en cada área se convierten a calificaciones estándar normalizadas basadas en características de edad y sexo, obtenidas de 200 niños caucásicos. También se dispone de un conjunto complementario de ilustraciones diseñadas específicamente para niños negros.

Test Aperceptivo de Relato de Cuentos para Niños (CAST). Esta prueba (pro.ed) se basa en la teoría adleriana y está diseñada para evaluar el funcionamiento emocional en niños de 6 a 13 años. El CAST consta de 31 ilustraciones a color acerca de las cuales los niños crean historias (Schneider, 1989; Schneider y Perney, 1990). Se elaboró para ser sensible a la raza y fue estandarizada en una muestra de 876 niños estadounidenses seleccionada como representativa. El CAST se califica en cuatro factores: adaptado, no adaptado, inmaduro y no investido. Los coeficientes de confiabilidad por consistencia interna y test-retest de las calificaciones de los facto-

res se encuentran entre .80 y .90. En el manual se presenta alguna evidencia sobre la validez de contenido, la validez con referencia a criterio y la validez de constructo, incluyendo perfiles de calificación para varios grupos clínicos de niños (con trastornos de déficit de atención, de conducta, de ansiedad, de negativismo y depresivos).

Test Aperceptivo de la Personalidad. El conjunto de ocho tarjetas de ilustraciones del Test Aperceptivo de la Personalidad (APT) (IDS) contrasta con el tono negativo o sombrío de las ilustraciones del TAT y de las escenas irreales representadas en ellas. Las ilustraciones del APT presentan a gente en escenarios familiares e incluyen a hombres y mujeres de diferentes edades y grupos étnicos (Karp, Holmstrom y Silber, 1990). Los examinados cuentan una historia acerca de cada ilustración y luego responden una serie de preguntas de opción múltiple diseñadas para proporcionar detalles adicionales acerca de las historias que son breves o crípticas. Las respuestas se califican en 16 medidas de personalidad: sentimientos hostiles activos; sentimientos hostiles pasivos; sentimientos hostiles totales; acciones hostiles activas; acciones hostiles pasivas; acciones hostiles totales; pasividad; dominio; confianza; estado de ánimo; imagen corporal; negación; degradación; inmadurez emocional; perspectiva, y distinción de carácter. También se dispone de una versión del APT Breve para Adultos y una versión para niños de 4 a 12 años (el Test Aperceptivo de Personalidad para Niños). Una prueba relacionada para niños y adultos con retraso mental (el Test Aperceptivo de Personalidad: Retraso Mental) consta de un conjunto de tarjetas con ilustraciones para abordar temas psicológicos de daño y aislamiento desde un grupo. El Test Aperceptivo de Personalidad se estandarizó en 517 hombres y 689 mujeres estudiantes no graduados de la universidad; 60 hombres y 71 mujeres adolescentes, y 20 hombres y 45 mujeres mayores, pero no se ha revisado ni se ha aplicado de manera amplia.

PROBLEMAS CON LAS TÉCNICAS PROYECTIVAS

Como se muestra en la breve revisión presentada en este capítulo, desde un punto de vista psicométrico estricto, las técnicas proyectivas dejan mucho que desear. La falta de objetividad en la calificación y la escasez de datos normativos representativos son en particular problemáticos para los especialistas en psicometría. No obstante, las críticas repetidas no parecen haber disminuido el entusiasmo de los psicólogos clínicos y los psiquiatras por las técnicas proyectivas. En la práctica se siguen usando con frecuencia la técnica de Rorschach, el TAT, el CAT, las frases incompletas y los dibujos proyectivos (vea Watkins, Campbell, Nieberding y Hallmark, 1995).

Parece que los clínicos, como muchos legos, consideran que las técnicas proyectivas poseen una especie de mística que las vuelve capaces de revelar la personalidad humana en mayor profundidad y detalle que los inventarios de personalidad, las escalas de calificación y las entrevistas controladas de manera más consciente y, por ende, más susceptibles al engaño. Cualesquiera que puedan ser las razones, Snyder (1974) encontró que la gente pone más fe, por lo regular, en las descripciones de la personalidad basadas en técnicas proyectivas que en las interpretaciones basadas en calificaciones de los inventarios de personalidad. Quizá las telenovelas, los asesinatos misteriosos, y otros programas de interés humano tratados en la televisión y los medios han vuelto a la gente más suspicaz y más inclinada a aceptar explicaciones complejas e intrincadas de conductas que a menudo pueden entenderse en términos de motivos humanos bastante ordinarios. Llama la atención que, cuando se les presentan descripciones agradables de sí mismos elaboradas supuestamente por astrólogos, incluso los escépticos comienzan a pensar: “después de todo, quizá haya algo en este negocio de la astrología” (Glick, Gottesman y Jolton, 1989).

PERSPECTIVAS PARA LA EVALUACIÓN DE LA PERSONALIDAD

Dado que este es el último de los cinco capítulos dedicados a la evaluación de la personalidad, así como el último capítulo del libro, parece adecuado efectuar una evaluación de dónde estamos y adónde podemos ir.

En comparación con la complejidad técnica de las pruebas de habilidad, las escalas de calificación de la personalidad, las listas de verificación y los inventarios, las pruebas proyectivas son relativamente imperfectas. El potencial valor práctico de una evaluación precisa de la personalidad es obvio, y el progreso reciente en el diseño psicométrico complejo de ciertos inventarios de características de personalidad e intereses sugiere que están por llegar progresos generales en la evaluación afectiva. Las aplicaciones de la teoría de la respuesta al ítem y de otras metodologías sofisticadas a la elaboración y calificación de las medidas de personalidad prometen mejoras en estos instrumentos. Las teorías y la investigación interesadas en los procesos cognoscitivos involucrados en los juicios clínicos también están haciendo contribuciones a la elaboración de métodos y productos de evaluación de la personalidad.

Hace una década y media, Ziskin (1986) mencionó una serie de signos del estado poco saludable de la psicología clínica: carencia de un sistema adecuado para clasificar los trastornos mentales, contaminación de los datos por efectos situacionales, evidencia de que las habilidades clínicas no mejoran con la práctica, dificultades para diferenciar entre la conducta normal y la psicopática, y problemas con la interpretación computarizada de los datos. Por desgracia, esos signos aún son visibles. Ziskin recomendaba tratar esta "enfermedad" con un enfoque más amplio hacia el uso de la computadora en el diagnóstico clínico, incluyendo no sólo al MMPI e instrumentos similares, sino también los datos demográficos y la información de entrevistas estructuradas. También anticipó que en el futuro se pondría mayor énfasis en las ventajas personales (como la buena apariencia), el estrés psicosocial y la entrevista estructurada. Por último, Ziskin recomendó a los clínicos tomar mayor conciencia del hecho de que existen límites en cuanto a lo que pueden descubrir acerca de la gente por medio de la evaluación psicológica.

Debido a su adaptabilidad a las circunstancias cambiantes y a la existencia de personas de gran capacidad que las apoyan, es poco probable que pruebas como el MMPI o la técnica de Rorschach sean reemplazadas pronto. Con seguridad se descubrirán nuevos usos para estas pruebas y tests más recientes obligarán a efectuar una reevaluación continua de esos y otros instrumentos clínicos.

Hace algunos años, Weiner (1983) predijo un futuro brillante para el psicodiagnóstico. Sin embargo, también advirtió que la realización de dicho futuro demanda "la investigación continua y cuidadosa sobre los métodos de psicodiagnóstico y la aplicación experta de los hallazgos de las pruebas psicológicas...". Éste sigue siendo un buen consejo. Casi una década después, Matarazzo (1992) siguió viendo un futuro brillante para el psicodiagnóstico clínico. Un ejemplo de este futuro es la elaboración de pruebas para identificar y analizar formas más específicas de psicopatología, como las reacciones de pánico y los trastornos depresivos. Matarazzo también pronosticó que se desarrollarían nuevas y mejores medidas de competencia personal para adaptarse a nuestro ambiente, incluyendo escalas de calidad de vida y conducta adaptada. Ahora, una década después, estamos empezando a notar la verificación de algunas de esas predicciones.

Aún más intrigante para Matarazzo que el futuro de la evaluación de la personalidad eran los desarrollos esperados en las medidas fisiológicas de la inteligencia, incluyendo índices derivados de potenciales cerebrales evocados e intrínsecos (EEG y PET) y la velocidad de conducción nerviosa. También se ha encontrado que los cambios en los potenciales cerebrales están asociados con calificaciones en medidas de rasgos de personalidad como introversión-extroversión.

Sea lo que fuere que pueda deparar el futuro para la evaluación clínica y de la personalidad, seguirá existiendo una necesidad continua de evaluar la efectividad de los instrumentos y procedimientos psicométricos en esas áreas. Entonces, como ahora, las preguntas principales concernientes a cualquier intento por evaluar la personalidad se concentrarán en la validez de los instrumentos y procedimientos para hacerlo. ¿En qué medida cumplen los métodos de evaluación los propósitos para los que fueron creados en la investigación, el diagnóstico clínico, la planeación del tratamiento y la evaluación de la efectividad de las intervenciones?

A diferencia de Nostradamus, los psicólogos no son conocidos por su habilidad para escudriñar el futuro. Aun así, todavía creo que esperan tiempos emocionantes y productivos a la examinación y evaluación psicológicas y a quienes han hecho de ésta la labor de su vida. La situación actual en este campo es muy diferente de lo que era al principio del siglo XX cuando se veía a la evaluación como la única manera en que los psicólogos podían trabajar fuera del salón de clase y el laboratorio.

Cualquier cosa que el siglo XXI, y no hablemos del tercer milenio, pueda deparar para el género humano, es lógico creer que los avances esperados en tecnología estarán acompañados por el progreso en la comprensión y ayuda a otra gente. La ciencia psicológica en general, y la evaluación psicológica en particular, deben apresurar el día en que realizarán sus contribuciones potenciales a esas metas.

RESUMEN

Las técnicas proyectivas son los instrumentos de evaluación de la personalidad menos estructurados. De manera tradicional, estas técnicas se han usado en contextos clínicos para identificar problemas personales y diagnosticar psicopatología. Sin embargo, algunas técnicas proyectivas han sido diseñadas o extendidas hacia el análisis de la personalidad normal.

Se han desarrollado varios tipos de técnicas proyectivas, incluyendo asociaciones de palabras, frases incompletas, dibujo de figuras humanas, respuestas a manchas de tinta e inventar historias a ilustraciones. Quienes defienden estas técnicas sostienen que los inventarios y otros instrumentos de autorreporte no logran llegar a las capas más profundas de la personalidad porque las personas no son conscientes de sus características y problemas o no los revelan. Como la calificación de las técnicas proyectivas regularmente es muy impresionista o subjetiva, se han encontrado dificultades para determinar la validez de esos instrumentos.

Las dos técnicas proyectivas más populares son la Técnica de Psicodiagnóstico de Rorschach y el Test de Apercepción Temática. También son dignos de atención el Estudio de Frustración Ilustrado de Rosenzweig, el Formulario de Frases Incompletas de Rotter, la Técnica de Manchas de Tinta de Holtzman y varias pruebas de historias con ilustraciones para niños, adultos mayores y grupos étnicos no blancos.

PREGUNTAS Y ACTIVIDADES

1. Compare las técnicas proyectivas con los inventarios de personalidad. Describa tanto los rasgos positivos como los negativos de ambos tipos de instrumentos y las condiciones bajo las cuales es más apropiado aplicar cada uno.

2. Elabore una lista alfabetizada de 25 sustantivos que correspondan a temas de interés para la gente de su grupo de edad: universidad, calificaciones, graduación, fracaso, sexo, matrimonio, religión,

madre, padre, carrera, salud, etc. Lea su lista a 12 conocidos y pida a cada quien que responda a cada palabra de la lista tan rápido como sea posible con la primera palabra que le venga a la mente. Registre el tiempo de la respuesta (en segundos) y la respuesta a cada palabra. Resuma los resultados en términos del número de respuestas de determinada clase dadas a cada palabra, los tiempos promedio de las respuestas y los conocimientos proporcionados sobre la personalidad de los sujetos. Revise la sección "Asociaciones de palabras" en este capítulo (páginas 413-414) antes de sacar cualquier conclusión.

3. Elabore diez frases incompletas que atañan a asuntos de interés para los estudiantes universitarios. Mecanografíelas (a doble espacio) fragmentadas en una hoja de papel y haga varias copias. Aplique su prueba de frases incompletas a un grupo de universitarios; pídales que completen cada fragmento de enunciado con una palabra o frase que tenga significado personal o que los refiera a un asunto de interés para ellos. Estudie las respuestas dadas y trate de analizarlas en términos de la personalidad de los sujetos. Escriba un informe donde resuma sus hallazgos.

4. Elabore dos historias incompletas que usted crea pueden revelar algo acerca de la personalidad del individuo que las complete. Ponga a prueba sus historias en varios de sus compañeros o amigos. ¿El contenido de las historias completadas sugiere algo significativo acerca de la personalidad del individuo? ¿Considera que esta técnica de evaluación de la personalidad es confiable y válida? ¿Por qué sí o por qué no?

5. Pida a varios individuos que hagan el dibujo de una persona en una hoja limpia de papel. Luego dígales que den vuelta a la hoja y hagan el dibujo de una persona del sexo opuesto. Recoja los dibujos y diga a los participantes que va a darlos a interpretar por un experto analista de la personalidad y que les dará a conocer más tarde esas interpretaciones. Algún tiempo después, presente la descripción de la personalidad dada en la página 323 a cada participante. Mezcle las oraciones en la interpretación de modo que el orden sea diferente para las distintas personas. Pida a cada persona que lea la descripción y le diga si es muy precisa, precisa, algo precisa, algo imprecisa, imprecisa o muy imprecisa como descripción de su personalidad. Tabule los resultados, interprételos tan bien como pueda y repórtelos a su maestro del curso. Después de haber completado el ejercicio, informe a los participantes del engaño y espere que reaccionen con buen humor a esta información.

6. Elabore la ilustración de una mancha de tinta colocando una gota grande de tinta negra en el medio de una hoja de papel blanco de $8^1/_2 \times 11$ pulgadas. Doble la hoja a la mitad de modo que la tinta quede en el interior y presiónela. Luego abra la hoja y deje que la mancha se seque. Repita el proceso con otras hojas de papel hasta que obtenga cinco manchas de tinta bastante detalladas y de preferencia simétricas. Luego aplique su prueba de manchas de tinta a varias personas. Pídales describir lo que ven en cada mancha, dónde lo ven en la mancha y qué aspecto de ésta (forma, color, textura u otra cualidad) les impulsó a dar esa respuesta. Registre las respuestas dadas por cada persona a cada mancha de tinta, y luego vea si puede decir algo acerca de las diversas personalidades a partir de sus respuestas a las cinco manchas de tinta. Compare los resultados con su conocimiento personal del individuo y con cualquier otro resultado de pruebas al que tenga acceso. Resuma sus hallazgos en un informe.

7. Busque en varias revistas populares y recorte o fotocopie cinco ilustraciones de personas en situaciones ambiguas. No debe ser inmediatamente obvio lo que están haciendo o pensando los sujetos de las fotografías. Presente éstas, una a la vez, a varias personas y pídales que cuenten una historia acerca de cada ilustración. Dígales que deben incluir en sus historias lo que está sucediendo, lo que llevó a ello (qué sucedió antes) y lo que resultará. Interprete las historias en términos de temas comunes, las acciones y sentimientos de los principales personajes, las presiones y frustraciones que ocurren en las historias, si las historias son por lo general placenteras o desagradables y si los fina-

les son optimistas o pesimistas (cómicos o trágicos). ¿Hay elementos comunes en todas las historias? ¿A partir de esas historias, puede usted decir algo acerca de la personalidad, las actitudes y los sentimientos del narrador?

8. Complete cada uno de los siguientes fragmentos de frase para mostrar sus verdaderos sentimientos.

1. Me gusta _____.
2. El mejor tiempo_____.
3. Mi madre _____.
4. Siento _____.
5. No puedo _____.
6. Otra gente_____.
7. Necesito _____.
8. Mi padre _____.
9. Esta universidad _____.
10. Deseo _____
11. No me gusta _____.
12. Soy muy _____.
13. Mis profesores _____.
14. Me preocupo acerca de _____.
15. Lamento que _____.

¿Qué temas consistentes, problemas y fuentes de conflicto son revelados en la forma en que completó esos enunciados? ¿Estuvo subjetivamente al tanto de cualquier tensión física o mental al responder a cualquiera de los reactivos? ¿Le tomó más tiempo completar algunos de los enunciados que otros? ¿Dio alguna respuesta extraña o inusual? De ser así, ¿qué podría significar? ¿Qué piensa de esta técnica de frases incompletas como método para analizar la personalidad? ¿Considera que revela algo detectable por un experto en diagnósticos psicológicos?

APÉNDICE A

ESTADÍSTICA DESCRIPTIVA

Cualquier clase de medición física (de tamaño, peso, coloración, etc.) que se realiza en seres vivos varía entre los miembros individuales de una especie. Los seres humanos presentan muchas diferencias físicas entre sí —en estatura, peso, presión sanguínea, agudeza visual, etc. Las diferencias individuales en esas variables físicas, además de en las capacidades cognoscitivas, rasgos de personalidad y conductas, son apreciables. Entre otras cosas, la gente difiere en sus capacidades, intereses, actitudes y temperamento. Algunas de esas diferencias individuales pueden medirse de manera más precisa que otras, como se refleja en el tipo de escala de medición.

ESCALAS DE MEDICIÓN

La medición de las variables físicas y psicológicas puede caracterizarse por el grado de refinamiento o precisión en términos de cuatro niveles o escalas: nominal, ordinal, de intervalo y de razón. Las medidas en una *escala nominal* sólo se utilizan para describir o nombrar, más que para indicar orden o magnitud. Algunos ejemplos de medición nominal son los números en los uniformes deportivos o las designaciones numéricas de variables demográficas como sexo (por ejemplo, hombre = 1, mujer = 2) y grupo étnico (blanco = 0, negro = 1, hispano = 3, asiático = 4). Dichos números son una forma conveniente de describir a individuos o grupos, pero no tiene sentido compararlos en términos de dirección o magnitud.

La medición en una *escala ordinal* es un poco más refinada que la medición nominal. Los números en una escala ordinal se refieren a las posiciones de objetos o acontecimientos en alguna variable. Por ejemplo, los números que designan el orden de terminación en una carrera u otra competencia están en una escala ordinal.

Un tercer nivel de medición es una *escala de intervalo*, en la cual diferencias numéricas iguales corresponden a diferencias iguales en cualquier característica medida. La escala de temperatura Celsius es un ejemplo de escala de intervalo. Así, la diferencia entre 40 °C y 60 °C es igual, en términos numéricos y de temperatura (calor), a la diferencia entre 10 °C y 30 °C. Las calificaciones estándar en las pruebas de inteligencia también se consideran mediciones de nivel de intervalo.

El nivel más alto, o más refinado, de medición es la *escala de razón*. Este tipo de escala tiene las características de una escala de intervalo así como un cero real: un valor de 0 en una escala de razón significa una completa ausencia de cualquier cosa que se esté midiendo. La medición realizada en una escala de razón permite que las razones numéricas se interpreten de manera significativa. Por ejemplo, la estatura se mide en una escala de razón. De modo que si Juan mide 1.82 y Pablo mide .91, es correcto decir que Juan es dos veces más alto que Pablo. Muchas variables físicas se miden en escalas de razón, pero la mayoría de las características psicológicas no son variables físicas. Las calificaciones en las pruebas psicológicas representan

medición en una escala ordinal o, cuando mucho, en una escala de intervalo. Por esta razón, aunque la calificación de Francisco en una prueba de inteligencia sea de 150 y la calificación de Jaime de 50, no podemos concluir que Francisco es tres veces más inteligente que Jaime. Pero si las calificaciones en la prueba son medidas de nivel de intervalo y Mary obtiene una calificación CI de 100, podemos decir que la diferencia en inteligencia entre Francisco y Mary (150-100) es igual a la diferencia entre Mary y Jaime (100-50).

DISTRIBUCIONES DE FRECUENCIA

El rango y la distribución de las diferencias individuales en las características físicas y mentales pueden representarse por medio de una distribución de frecuencia de las calificaciones obtenidas en una prueba o algún otro instrumento psicométrico. En su forma más simple, una *distribución de frecuencia* es una lista integrada por las posibles calificaciones y la cantidad de personas que obtuvo cada calificación. Suponga que en una prueba de cinco reactivos se otorga un punto por cada respuesta correcta. Entonces, las posibles calificaciones son 0, 1, 2, 3, 4 y 5. Si 25 personas presentan la prueba, la distribución de frecuencia de sus calificaciones podría verse de la siguiente manera:

CALIFICACIÓN	FRECUENCIA
5	1
4	4
3	9
2	6
1	3
0	2

Advierta que dos personas dieron respuestas incorrectas a los cinco reactivos, nueve respondieron bien a tres reactivos, y una persona respondió los cinco reactivos de manera correcta.

Intervalos de calificación

Cuando el rango de calificaciones en una prueba es amplio, digamos 25 puntos o más, puede ser conveniente agrupar las calificaciones en intervalos. Como ejemplo de ello, tenemos que las calificaciones del cociente de inteligencia (CI) en la Escala de Inteligencia para Adultos de Wechsler (WAIS) fluctúan entre 43 y 152 aproximadamente. Los cálculos efectuados sobre esas calificaciones pueden simplificarse al agruparlas en intervalos de 5 puntos CI, comenzando con el intervalo 43-47 y contando hasta el intervalo 148-152 (vea la columna 1 de la tabla A.1). Esto nos da 22 intervalos en lugar de los 110 intervalos (CI de 43 a 152) que podrían resultar si se asignara un intervalo para cada calificación posible. La utilización del menor número de intervalos tiene poco efecto sobre la precisión estadística que se calcula a partir de la distribución de frecuencia de las calificaciones CI de la WAIS, y es una manera más eficiente de describir las calificaciones.

TABLA A.1 Distribución de frecuencia de los CI de escala completa en la escala WAIS*

INTERVALO DE CI	NÚMERO DE EXAMINADOS (FRECUENCIA)	INTERVALO DE CI	NÚMERO DE EXAMINADOS (FRECUENCIA)
93–97	255	148–152	1
88–92	220	143–147	0
83–87	135	138–142	3
78–82	107	133–137	12
73–77	55	128–132	26
68–72	49	123–127	64
63–67	18	118–122	145
58–62	11	113–117	165
53–57	6	108–112	224
48–52	3	103–107	274
43–47	1	98–102	278

*Datos de D. Wechsler, *The Measurement and Appraisal of Adult Intelligence,* 4ª edición. Baltimore: Williams & Wilkins, 1958, p. 253.

Histograma y polígono de frecuencia

Una distribución de frecuencia de calificaciones puede representarse gráficamente como un histograma o un polígono de frecuencia. Para elaborar un *histograma*, primero deben determinarse los límites exactos de los intervalos de calificación. Los *límites exactos* de un intervalo se calculan restando .5 del límite inferior y sumando .5 al límite superior del intervalo. Por ejemplo, los límites exactos del intervalo 43-47 en la tabla A.1 son 42.5 y 47.5, y los límites exactos del intervalo 148-152 son 147.5 y 152.5. Después de que se han calculado los límites exactos de todos los intervalos, la frecuencia correspondiente a cada intervalo se representa como una barra vertical con una anchura que se extiende sobre los límites exactos y una altura proporcional al número de calificaciones que cae en el intervalo. La figura A.1 es un histograma de la distribución de frecuencia de la tabla A.1.

Una distribución de frecuencia también puede ser representada por una serie de segmentos de línea conectados. En la figura A.2, los puntos que corresponden a las frecuencias y los puntos medios de los intervalos de calificación en la tabla A.1 se unieron para formar un *polígono de frecuencia*.

La curva normal

El polígono de frecuencia que muestra la figura A.2 no es una curva suave, pero su apariencia es similar a la de una curva simétrica en forma de campana. Más personas obtuvieron calificaciones aproximadas de 100 (en realidad, 98-102) que cualquier otra calificación, y cada vez menos personas obtuvieron calificaciones menores o mayores de 100. Si el polígono de frecuencia fuera perfectamente simétrico, suave y con forma de campana, se parecería a la figura A.3.

La gráfica mostrada en la figura A.3, la cual puede ser descrita por una ecuación matemática, se denomina *curva normal*. Las calificaciones en el eje base de esta curva son *calificaciones estándar* (*calificaciones z*), cuyo cálculo se describe en el capítulo 4. Esas calificaciones *z* sirven como un método estándar conveniente de expresar y comparar las calificaciones de la

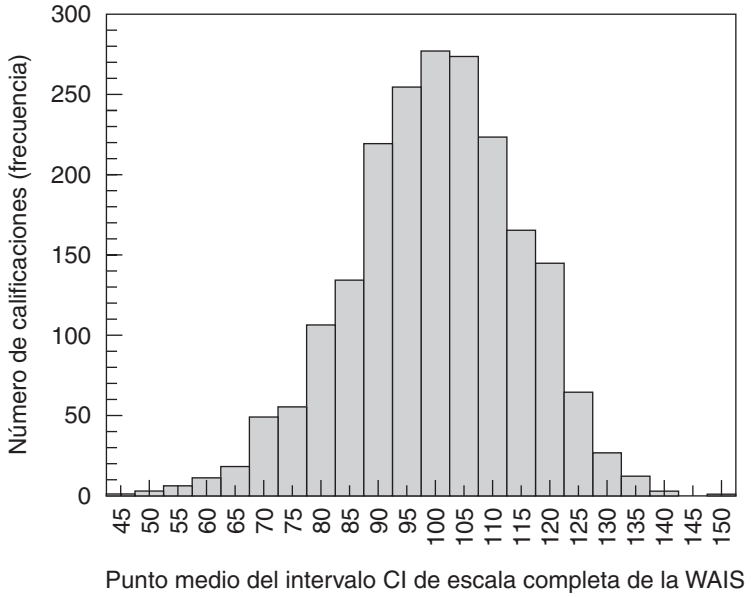


FIGURA A.1 Histograma de la distribución de frecuencia de la tabla A.1

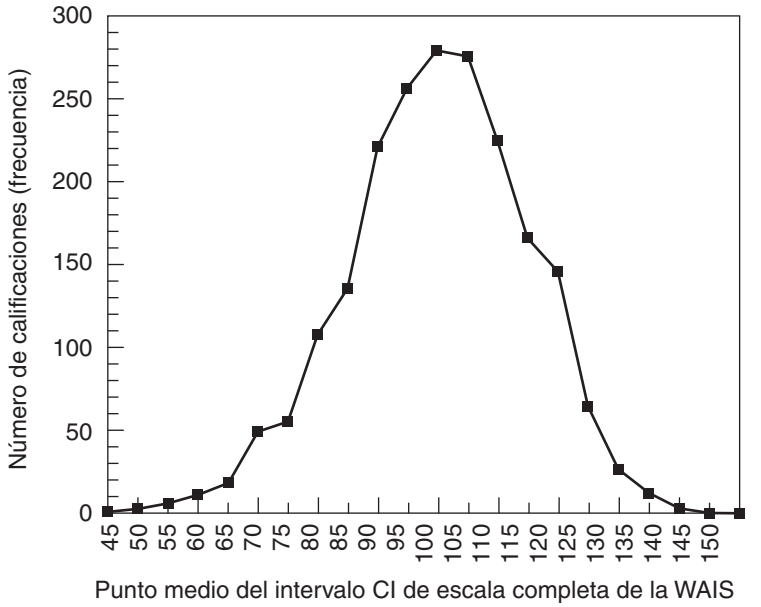


FIGURA A.2 Polígono de frecuencia de la distribución de frecuencia de la tabla A.1

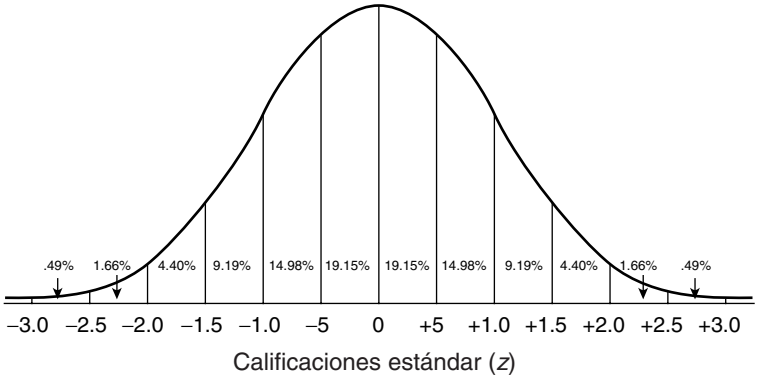


FIGURA A.3 Distribución normal estándar

misma persona en dos o más pruebas, o las calificaciones de dos o más personas en la misma prueba.

Cierto porcentaje del área situada bajo la curva en la figura A.3 cae entre dos calificaciones z cualesquiera. Este porcentaje puede corresponder al porcentaje de un grupo de personas cuyas calificaciones crudas en la prueba, al convertirse a calificaciones z, caen dentro del rango de las dos calificaciones z. Por ejemplo, 19.15% del área bajo la curva y, en consecuencia, 19.15% de una distribución normal de calificaciones en la prueba caen entre $z = 0$ y $z = .5$ (o $z = 0$ y $z = -.5$). Por otro lado, sólo 1.66% del área bajo una curva normal cae entre $z = +2.0$ y $z = +2.5$ (o $z = -2.0$ y $z = -2.5$).

El rango teórico de calificaciones z en una distribución normal es menos infinito ($-\infty$) a más infinito ($+\infty$), pero más de 99% del área bajo la curva normal (o 99% de una distribución normal de calificaciones en la prueba) cae entre las calificaciones z de -3.00 y $+3.00$. Por supuesto, cuando se convierte la calificación cruda de una prueba a una calificación z, el resultado no siempre es una de las 13 calificaciones z presentadas en el eje horizontal de la figura A.3. Debe utilizarse una tabla especial como la que se presenta en el apéndice B de este libro o un programa de computadora para determinar el porcentaje del área que cae debajo de, y por sustracción entre, dos valores cualesquiera de z.

A finales del siglo XIX y principios del XX, hubo mucha especulación concerniente a si la curva normal expresaba una ley inherente de la naturaleza. La razón para esta creencia fue que las distribuciones de frecuencia de las mediciones efectuadas sobre muchas características de los organismos vivos son de una forma normal aproximada. En efecto, buena parte de la teoría matemática de inferencia estadística, la cual es muy importante en la investigación psicológica y educativa, se basa en la suposición de una distribución normal de las mediciones. Sin embargo, debemos ser cuidadosos y no glorificar la curva normal. Aunque muchas pruebas están elaboradas de tal forma que sus calificaciones se distribuyen de manera aproximadamente normal, las distribuciones de frecuencia de las calificaciones de otras pruebas son muy asimétricas o sesgadas. Una situación común es una distribución de calificaciones con *sesgo positivo* (pocas calificaciones altas y muchas calificaciones bajas), la cual representa los resultados de una prueba que quizá era demasiado difícil (vea la figura A.4). Menos común es una distribución con *sesgo negativo* (muchas calificaciones altas y pocas calificaciones bajas), la cual ocurre cuando una prueba es demasiado sencilla.

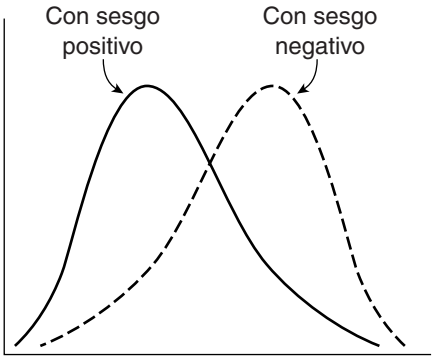


FIGURA A.4 Distribuciones de frecuencia sesgadas

MEDIDAS DE TENDENCIA CENTRAL

Además de graficar la distribución de un conjunto de calificaciones de prueba, es conveniente tener alguna medida de la calificación típica o promedio. Es posible calcular tres tipos de promedio: moda, mediana y media aritmética.

Moda

La *moda* de un conjunto de calificaciones de prueba es la calificación obtenida por el mayor número de personas. En la prueba de cinco reactivos mencionada antes, más personas (9) obtuvieron una calificación de 3 que cualquier otra calificación, por lo que la moda es 3. Cuando las calificaciones de la prueba se agrupan en intervalos, la moda es el punto medio del intervalo que contiene el mayor número de calificaciones. En la tabla A. 1, el intervalo de calificación 98–102 contiene el mayor número de calificaciones (278); el punto medio de ese intervalo— $(98 + 102)/2 = 100$— es la moda de esa distribución de frecuencia.

Como se observa en el polígono de frecuencia de la figura A.2, la moda es la calificación que corresponde al punto más alto en una distribución de frecuencia. La figura A.2 ilustra una distribución *unimodal* con un solo pico. En ocasiones una distribución de frecuencia tiene más de un pico; es *bimodal* si tiene dos picos y *multimodal* cuando presenta más de dos.

Mediana

La *mediana* (Mdn) de un conjunto de calificaciones es la calificación intermedia, es decir, la calificación por debajo de la cual cae la mitad de las calificaciones. La mediana de 7, 6, 9, 5 y 3 es 6, porque 6 se encuentra en el centro cuando esas cinco calificaciones se ordenan de la más alta a la más baja. Cuando el número de calificaciones es par, la mediana se define como la media de las dos calificaciones intermedias.

Se requieren unos cuantos pasos para calcular la mediana de una distribución de frecuencia, pero puede encontrarse con bastante rapidez interpolando dentro del intervalo en el que cae. Para ilustrar el procedimiento, se calculará la mediana de la distribución de frecuencia de la tabla A.1. El número total de calificaciones es 2,052, así que la mediana es la calificación CI por debajo y por encima de la cual caen $.5(2{,}052) = 1{,}026$ calificaciones. Al sumar sucesivamente

las frecuencias en la columna 2 de la tabla A.1, encontramos que hay 860 calificaciones hasta el intervalo 93-97 y 1,138 hasta el intervalo 98-102. Expresado en términos de los límites superiores exactos de los intervalos, decimos que 860 calificaciones caen por debajo de 97.5 y 1,138 calificaciones caen por debajo de 102.5. Como la mediana es la calificación por debajo de la cual caen 1,026 calificaciones, se encuentra entre 97.5 y 102.5. Para encontrar la mediana exacta, formamos la razón (Mdn $-97.5)/(102.5 - 97.5) = (1{,}026 - 860)/(1{,}138 - 860)$. Al resolver esta ecuación obtenemos un valor de 100.49 para la mediana.

El procedimiento descrito líneas arriba para encontrar la mediana de una distribución de frecuencia puede simplificarse como:

$$\text{Mdn} = L + \frac{w(.5n_t - n_b)}{n_i} \qquad \textbf{(A.1)}$$

En esta fórmula, L es el límite exacto inferior y w es la amplitud del intervalo que contiene la mediana, n_t es el número total de calificaciones en la distribución, n_b es el número de calificaciones que caen por debajo del intervalo que contiene la mediana, y n_i es el número de calificaciones que caen en el intervalo que contiene la mediana.

El operador de sumatoria

Antes de considerar el procedimiento para calcular la media aritmética, debe familiarizarse con el símbolo especial Σ. Este símbolo, la letra mayúscula griega sigma, es una forma abreviada de designar la operación matemática de la suma. De este modo,

$$\sum X = X_1 + X_2 + X_3 + \ldots + X_n.$$

A manera de ejemplo, considere las tres calificaciones $X_1 = 2$, $X_2 = 4$ y $X_3 = 1$. La suma de esas calificaciones es $\Sigma X = 2 + 4 + 1 = 7$. De manera similar, la suma de cuadrados de las tres calificaciones es

$$\sum X^2 = X_1^2 + X_2^2 + X_3^2 = 2^2 + 4^2 + 1^2 = 4 + 16 + 1 = 21.$$

La suma de los productos de las dos variables, X y Y, se expresa como

$$\sum XY = X_1Y_1 + X_2Y_2 + X_3Y_3.$$

Si $Y_1 = 3$, $Y_2 = 5$, $Y_3 = 2$ y los valores de X son los mismos que en el problema precedente,

$$\sum XY = 2(3) + 4(5) + 1(2) = 6 + 20 + 2 = 28.$$

Media aritmética

Si bien la moda es fácil de calcular, se ve muy afectada por la forma de la distribución de frecuencia de calificaciones. La mediana, que se ve menos afectada por la forma de la distribución de frecuencia, es la medida de tendencia central preferida cuando la distribución es muy asimétrica o sesgada. Como resulta engorroso trabajar con la mediana en la teoría estadística, la media aritmética es la medida de tendencia central (promedio) más popular. La media aritmética de un

conjunto de calificaciones (X) se determina sumando las calificaciones y dividiendo la suma resultante entre el número de calificaciones:

$$\overline{X} = \sum X/\text{n}. \qquad \textbf{(A.2)}$$

La media de las calificaciones X en el problema precedente es 7/3 = 2.33.

Cuando las calificaciones se agrupan en la forma de una distribución de frecuencia, la media puede encontrarse con mayor rapidez (1) multiplicando el punto medio (X_i') de cada intervalo por la frecuencia (f_i) en el intervalo; (2) sumando esos productos fX', y (3) dividiendo la suma resultante de los productos entre el número total de calificaciones (n):

$$\overline{X} = \sum X'/\text{n}. \qquad \textbf{(A.3)}$$

Por ejemplo, la media aritmética del problema de cinco reactivos descrito antes (página 429) se calcula de la siguiente manera:

$$\frac{2(0) + 3(1) + 6(2) + 9(3) + 4(4) + 1(5)}{25} = \frac{63}{25} = 2.52.$$

A manera de ejercicio, verifique si la media aritmética de la distribución de frecuencia dada en la tabla A.1 es 99.96.

PERCENTILES, DECILES Y CUARTILES

La mediana se conoce en ocasiones como el percentil 50$^{o.}$ porque 50% de las calificaciones cae por debajo de ella. Una distribución de frecuencia puede dividirse en 100 percentiles; el *percentil p* es el valor por debajo del cual cae el p por ciento de las calificaciones. Por ejemplo, el percentil 25$^{o.}$ es el valor por debajo del cual cae 25% de las calificaciones, y el percentil 75$^{o.}$ es el valor por debajo del cual cae 75% de las calificaciones. Cualquier percentil puede calcularse por un procedimiento similar al descrito antes para encontrar la mediana.

Además de los percentiles, una distribución de frecuencia puede dividirse en décimos (*deciles*), quintos (*quintiles*) o cuartos (*cuartiles*). El cuarto decil (o percentil 40$^{o.}$) es el valor por debajo del cual caen cuatro décimos de las calificaciones, y el tercer cuartil (o percentil 75$^{o.}$) es el valor por debajo del cual caen tres cuartos de las calificaciones. Note que el percentil 50$^{o.}$, el quinto decil y el segundo cuartil son iguales al mismo valor numérico.

MEDIDAS DE VARIABILIDAD

Una medida del promedio o tendencia central no proporciona, por sí misma, una descripción analítica adecuada de una muestra de calificaciones. Las distribuciones de frecuencia de las calificaciones difieren no sólo en sus promedios, sino también en su grado de variabilidad (dispersión), simetría y carácter puntiagudo. Se describirán tres medidas de variabilidad: el rango, el rango semiintercuartilar y la desviación estándar.

Rango y rango semiintercuartilar

El *rango* simple, definido como la calificación más alta menos la calificación más baja, es la medida de variabilidad más sencilla de calcular.[1] El rango de calificaciones en el problema de cinco reactivos descrito antes es $5 - 0 = 5$, y el rango de las calificaciones CI en la tabla A.1 es $152 - 43 = 109$. En la mayoría de los casos, el rango es una medida pobre de variabilidad debido a que se ve muy afectado por una sola calificación muy alta o muy baja. Un tipo modificado de rango conocido como *rango semiintercuartilar* se utiliza en ocasiones como un índice de variabilidad cuando la distribución de calificaciones es muy sesgada. El rango semiintercuartilar, o Q, se calcula como la mitad de la diferencia entre el percentil 75º. (tercer cuartil) y el percentil 25º. (primer cuartil).

Como ejercicio, verifique si, para la distribución de frecuencia dada en la tabla A.1, el primer cuartil es 90.41, el tercer cuartil es 110.33 y el rango semiintercuartilar es 9.96. Los dos cuartiles pueden encontrarse por el mismo tipo de procedimiento de interpolación lineal que se empleó para calcular la mediana. Como el primer cuartil es la calificación por debajo de la cual caen $.25(2{,}052) = 513$ calificaciones, interpolamos dentro del intervalo 87.5 a 92.5. Luego se determina el primer cuartil resolviendo para Q_1 en la expresión:

$$\frac{Q_1 - 87.5}{92.5 - 87.5} = \frac{513 - 385}{605 - 385}.$$

Para encontrar el tercer cuartil, el cual es la calificación $.75(2{,}052) = 1{,}539$º., interpolamos dentro del intervalo 107.5 a 112.5. Luego se calcula el tercer cuartil resolviendo para Q_3 en la expresión:

$$\frac{Q_3 - 107.5}{112.5 - 107.5} = \frac{1539 - 1412}{1636 - 1412}.$$

Desviación estándar

La medida más común de variabilidad, la *desviación estándar*, resulta apropiada cuando la media aritmética es el promedio reportado. La desviación estándar de una muestra de calificaciones puede calcularse a partir de:

$$s = \sqrt{\frac{\Sigma X^2 - (\Sigma X)^2/n}{n - 1}} \qquad \textbf{(A.4)}$$

Por ejemplo, para encontrar la desviación estándar de 7, 6, 9, 5 y 3, comenzamos por calcular $\Sigma X = 30$ y $\Sigma X^2 = 200$. Entonces,

$$\frac{\Sigma X^2 - (\Sigma X)^2/n}{n - 1} = \frac{200 - 30^2/5}{4} = 5.$$

[1]Éste es el *rango exclusivo*. El *rango inclusivo* es igual al rango exclusivo más 1.

la cual es la *varianza* de nuestros cinco números. Al extraer la raíz cuadrada de la varianza se obtiene 2.24, la desviación estándar.

Al establecer que $\Sigma X = \Sigma f(X')$ y $\Sigma X^2 = \Sigma f(X'^2)$, donde f es la frecuencia y X' el punto medio de un intervalo, podemos usar la fórmula A.4 para calcular la desviación estándar de una distribución de frecuencia. A manera de ejercicio, confirme que la desviación estándar de la distribución de frecuencia del problema de cinco reactivos referido antes es 1.26, y que la desviación estándar de la distribución de frecuencia dada en la tabla A.1 es 14.85.

CORRELACIÓN Y REGRESIÓN LINEAL SIMPLE

El método de correlación ha sido empleado con frecuencia en el análisis de los datos de prueba, y también es muy importante en la teoría clásica de las pruebas. La correlación se ocupa de determinar el grado en que dos conjuntos de medidas, como las calificaciones en una prueba de inteligencia y las notas escolares, están relacionadas. La magnitud y dirección de la relación entre dos variables se expresa como un índice numérico, al cual se le conoce como el *coeficiente de correlación*. De los muchos tipos diferentes de coeficiente de correlación, el más popular es el *coeficiente producto-momento* de Pearson, o *r*. Su valor fluctúa entre −1.00 (una relación inversa perfecta) y +1.00 (una relación directa perfecta). Sin embargo, el coeficiente *r* de Pearson no es el único coeficiente de correlación que se utiliza para analizar y aplicar las calificaciones de prueba. Por ejemplo, el coeficiente de correlación biserial puntual, el cual se describe en el capítulo 4, se usa ampliamente en el análisis de reactivos.

Cálculo del coeficiente producto-momento

La tabla A.2 ilustra los cálculos iniciales para determinar el coeficiente de correlación entre 30 pares de calificaciones $X-Y$. Dejemos que X sea una calificación en una prueba de capacidad y Y una calificación del desempeño en el trabajo. De este modo, la primera persona obtuvo 44 en la prueba de capacidad y 69 en la calificación del desempeño, mientras que las calificaciones correspondientes para la segunda persona son 38 y 46. Los encabezados de las columnas indican los pasos para calcular *r*:

1. Calcule X^2, Y^2 y XY para cada persona (columnas 4, 5 y 6).
2. Sume las columnas X, Y, X^2, Y^2 y XY (columnas 2 a 6) y sustituya esas sumas en la siguiente fórmula:

$$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}. \qquad \textbf{(A.5)}$$

Como $\Sigma X = 1{,}498$, $\Sigma Y = 1{,}511$, $\Sigma X^2 = 79{,}844$, $\Sigma Y^2 = 79{,}641$ y $\Sigma XY = 77{,}664$,

$$r = \frac{30(77{,}664) - (1498)(1511)}{\sqrt{[30(79{,}844) - (1498)^2][30(79{,}641) - (1511)^2]}} = .524.$$

El significado de la correlación

El método de correlación es útil en el campo de las pruebas psicológicas por diversas razones, entre las cuales está el hecho de que la correlación implica la posibilidad de hacer predicciones.

TABLA A.2 Cálculo de las sumas para determinar la correlación producto-momento

(1) PERSONA	(2) X	(3) Y	(4) X^2	(5) Y^2	(6) XY
1	44	69	1,936	4,761	3,036
2	38	46	1,444	2,116	1,748
3	56	51	3,136	2,601	2,856
4	54	44	2,916	1,936	2,376
5	66	53	4,356	2,809	3,498
6	52	49	2,704	2,401	2,548
7	46	43	2,116	1,849	1,978
8	36	35	1,296	1,225	1,260
9	44	37	1,936	1,369	1,628
10	60	69	3,600	4,761	4,140
11	22	31	484	961	682
12	72	47	5,184	2,209	3,384
13	56	45	3,136	2,025	2,520
14	52	41	2,704	1,681	2,132
15	50	39	2,500	1,521	1,950
16	64	65	4,096	4,225	4,160
17	40	36	1,600	1,296	1,440
18	28	59	784	3,481	1,652
19	68	70	4,624	4,900	4,760
20	48	53	2,304	2,809	2,544
21	32	51	1,024	2,601	1,632
22	74	63	5,476	3,969	4,662
23	42	54	1,764	2,916	2,268
24	50	52	2,500	2,704	2,600
25	40	49	1,600	2,401	1,960
26	58	48	3,364	2,304	2,784
27	62	60	3,844	3,600	3,720
28	54	64	2,916	4,096	3,456
29	60	55	3,600	3,025	3,300
30	30	33	900	1,089	990
Sumas:	1,498	1,511	79,844	79,641	77,664

La precisión con que puede predecirse la calificación de una persona en la medida Y a partir de su calificación en la medida X depende de la magnitud de la correlación entre las calificaciones en las dos variables. Entre más cercano sea el coeficiente de correlación a un valor absoluto de 1.00 (sea $+1.00$ o -1.00), menor será el error promedio cometido al predecir las calificaciones Y a partir de las calificaciones X. Por ejemplo, si la correlación entre las pruebas X y Y es cercana a $+1.00$, puede predecirse con confianza que una persona que obtenga una calificación alta en la variable X también la obtendrá en la variable Y, y quien obtenga una calificación baja en X obtendrá una calificación baja en Y. Por otro lado, si la correlación es cercana a -1.00, puede predecirse con cierta confianza que una persona que obtenga una calificación elevada en X ob-

tendrá una calificación baja en *Y*, y quien califique bajo en *X* calificará alto en *Y*. Entre más cercano sea el valor de *r* a +1.00 o a −1.00, más precisa será la predicción; entre más cercano sea *r* a .00, menos precisa será la predicción. Cuando *r* = .00, la predicción de la calificación de una persona en una variable a partir de su calificación en la otra variable no será mejor que el azar.

Es importante recordar que la correlación implica predicción, pero no supone causalidad. El hecho de que dos variables estén relacionadas no significa que una variable sea por necesidad causa de la otra. Ambas variables pueden estar influidas por una tercera variable, y la correlación entre las dos primeras es un reflejo de esta causa común. Por ejemplo, puede demostrarse que las edades mentales de un grupo de niños con una amplia gama de edades cronológicas se correlacionan de manera positiva con el tamaño de sus zapatos. La edad mental no es causa del tamaño de los zapatos ni viceversa, sino que la correlación positiva entre esas dos variables se debe a la influencia de una tercera variable, madurez o crecimiento físico, sobre la edad mental y el tamaño de los zapatos. El hecho de que dos variables tengan una correlación significativa facilita la predicción del desempeño en una a partir del desempeño en la otra, pero no proporciona información directa sobre si las dos variables mantienen una conexión causal.

Regresión lineal simple

El coeficiente de correlación producto-momento, el cual es una medida de la relación *lineal* entre dos variables, es en realidad un coproducto del procedimiento estadístico para encontrar la ecuación de la línea recta que mejor se ajusta al conjunto de puntos que representan los valores pareados *X*-*Y*. Para ilustrar el significado de esta afirmación, los pares de valores *X*-*Y* presentados en la tabla A.2 se trazan como un *diagrama de dispersión* en la figura A.5. Es claro que los

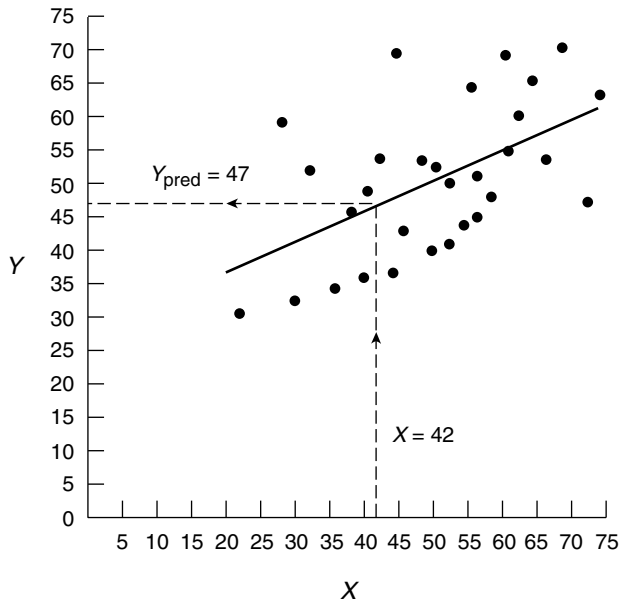


FIGURA A.5 Diagrama de dispersión, de las calificaciones dadas en la tabla A.2, que muestra la línea de regresión y un problema ilustrativo

30 puntos no caen en la misma línea recta, pero puede ajustarse una línea a través de los puntos de tal manera que la suma de las distancias verticales elevadas al cuadrado de los puntos a partir de la línea sea lo más pequeña posible. Una fórmula para encontrar esta *línea de regresión de cuadrados mínimos* es:

$$Y_{\text{pred}} = r\frac{s_y}{s_x}(X - \overline{X}) + \overline{Y}, \qquad \textbf{(A.6)}$$

donde $\overline{X}$ y $\overline{Y}$ son las medias y s_x y s_y las desviaciones estándar de las variables *X* y *Y*. Para los datos en la tabla A.2, $\overline{X} = 49.93$, $\overline{Y} = 50.37$, $s_x = 13.19$, $s_y = 11.04$ y $r = .52$. Al ingresar esos números en la fórmula A.6 y simplificar se obtiene la ecuación lineal $Y_{\text{pred}} = .44X + 28.64$. Al usar esta ecuación puede predecirse la calificación de una persona en la variable *Y* con una precisión mejor que el azar a partir de su calificación en la variable *X*. Por ejemplo, como lo ilustran las líneas punteadas en la figura A.5, si $X = 42$, $Y_{\text{pred}} = .44(42) + 28.64 = 47.12$. Esto significa que si una persona obtiene una calificación de 42 en la variable *X*, la mejor estimación de su calificación en la variable *Y* es aproximadamente 47.

REGRESIÓN MÚLTIPLE Y ANÁLISIS FACTORIAL

Muchos otros procedimientos estadísticos se emplean al analizar las calificaciones de prueba y se utilizan con propósitos de evaluación y predicción. Entre esos procedimientos se encuentran el análisis de regresión múltiple, el análisis discriminante, el análisis de perfiles, el escalamiento multidimensional y el análisis factorial. Todos esos temas se consideran en detalle en libros de estadística y psicometría avanzadas (por ejemplo, Nunnally y Bernstein, 1994), por lo que en interés del espacio aquí sólo se considerarán el primero y el último.

Regresión múltiple

El análisis de regresión lineal simple que involucra una variable independiente (*X*) puede ser extendido a dos o más variables independientes. Dejemos que *Y* represente una variable criterio como el aprovechamiento académico o el desempeño en el trabajo, y hagamos que $X_1, X_2, X_3 ..., X_n$ representen una serie de *n* variables independientes (predictoras). En símbolos, el problema de predecir el criterio a partir de esas variables puede expresarse como el encontrar una solución a la ecuación

$$Y_p = B_1X_1 + B_2X_2 + B_3X_2 + ... + B_nX_n + A.$$

En esta ecuación, Y_p es el valor anticipado de *Y*, las *B* son los pesos de regresión no estandarizados (calificación cruda) para las variables independientes correspondientes y *A* es una constante que denota el punto en el cual el plano de regresión *n*-dimensional intersecta con el eje *Y*. La importancia relativa o significancia de las *n* variables independientes para predecir el criterio está indicada por la magnitud de los coeficientes de regresión estandarizados (β), los cuales son iguales a

$\beta_i = (s_i/s_y)B_i$. Un índice de la precisión combinada de las variables independientes para predecir el criterio es R, el coeficiente de correlación múltiple.

Como ejemplo práctico, suponga que un investigador está interesado en realizar un análisis de regresión múltiple para determinar la efectividad de las calificaciones de un grupo de estudiantes de primer grado en tres variables independientes [una prueba de preparación para la lectura (X_1), una prueba de inteligencia (X_2), y un entero (1 = varón, 2 = mujer) que indica el sexo (género) del niño (X_3)] para predecir las calificaciones (Y_p) en una prueba de aprovechamiento en lectura aplicada a los estudiantes al terminar el primer grado. Los cálculos requeridos para este análisis de regresión múltiple pueden realizarse con facilidad por una computadora usando un paquete estadístico como el SPSS.[2]

La entrada a la computadora para el problema pueden ser las calificaciones de los estudiantes en las tres pruebas y el valor codificado para el sexo, o, de manera alternativa, las correlaciones entre las cuatro variables. Supongamos que las correlaciones entre las cuatro variables ya han sido calculadas. Tomando nota de que los subíndices 1, 2 y 3 se refieren a las variables independientes 1, 2 y 3 y que el subíndice y se refiere a la variable dependiente, esas correlaciones son $r_{12} = .466$, $r_{13} = .055$, $r_{23} = .072$, $r_{y1} = .612$, $r_{y2} = .541$ y $r_{y3} = .197$. Además de las correlaciones entre las variables necesitamos ingresar sus medias y desviaciones estándar a la computadora. Éstas son $\overline{X}_1 = 49.0$, $\overline{X}_2 = 102.8$, $\overline{X}_3 = 1.48$, $\overline{Y} = 26.0$, $s_1 = 10.3$, $s_2 = 14.7$, $s_3 = .501$ y $s_y = 8.10$.[3]

Un programa de regresión múltiple típico calcula, junto con otros estadísticos, los pesos de regresión estandarizados (β) y no estandarizados (B) para las variables independientes, el intercepto Y, el coeficiente de correlación múltiple (R), los errores estándar de los pesos de regresión, y las razones críticas para determinar la significancia estadística de los pesos de regresión. Para el problema anterior, los pesos de beta (β) son $\beta_1 = .4556$, $\beta_2 = .3179$, y $\beta_3 = .1490$; los pesos de B son $B_1 = .3583$, $B_2 = .1752$, y $B_3 = 2.4098$; y el intercepto Y (A) es $= -13.1338$. Así, la ecuación de regresión de la calificación cruda (no estandarizada) es

$$Y_p = .3583X_1 + .1752X_2 + 2.4098X_3 - 13.1338$$

Las pruebas estadísticas realizadas en los pesos de regresión indican que todos son significativos, con la primera variable independiente siendo el mejor pronosticador seguido de la segunda variable independiente. La efectividad global de los tres pronosticadores, en combinación, para predecir las calificaciones en la variable criterio está indicada por un coeficiente de correlación múltiple de $R = .693$, un valor significativamente alto.

Existen muchos otros aspectos en el análisis de regresión y se han escrito libros enteros sobre el tema. Este breve tratamiento, el cual apenas ha tocado la superficie de una importante técnica estadística en psicometría, debe servir para estimular el interés del lector por conocer tratamientos más amplios (por ejemplo, Kleinbaum, Kupper, Muller y Nizam, 1998).

[2]Los cálculos requeridos también pueden hacerse usando el programa A-5 “Computer Programs for Psychological Testing and Assessment”.

[3]Esos valores fueron tomados de un problema descrito en las páginas 137 y 138 de Glass y Hopkins (1996).

Análisis factorial

El propósito principal del *análisis factorial* es reducir el número de variables en un grupo de medidas tomando en cuenta el traslape (correlaciones) entre ellas. En el campo de las pruebas psicológicas, el problema es encontrar unos cuantos factores sobresalientes que expliquen la mayor parte de la varianza de un grupo de calificaciones en diferentes pruebas. La gran variedad de procedimientos existentes para extraer esos factores de las calificaciones de prueba se basan en un teorema particular: la varianza observada (total) de una prueba (s^2_{obs}) es igual a la suma de la varianza debida a factores que la prueba tiene en común con otras pruebas (s^2_{com}), la varianza específica para la prueba misma (s^2_{esp}), y la varianza producida por errores de medición (s^2_{err}). En consecuencia, la fórmula 5.1 del capítulo 5 (página 86) puede replantearse de la siguiente manera:

$$s^2_{obs} = s^2_{com} + s^2_{esp} + s^2_{err}. \quad \textbf{(A.7)}$$

En la fórmula A.7, a la cual se hizo referencia en el capítulo 5 como la varianza real de una prueba (s^2_{rea}), la varianza observada se divide en varianza de factores comunes y varianza de factores específicos. La parte de la varianza observada que se debe a factores comunes se denomina *comunalidad* de una prueba, mientras que la parte debida a factores específicos es su *especificidad*. A partir de estas definiciones y de las fórmulas 5.2 y A.7, podemos expresar la ecuación:

$$\text{confiabilidad} = \text{comunalidad} + \text{especificidad}. \quad \textbf{(A.8)}$$

Un componente de esta ecuación, la comunalidad de una prueba, se obtiene de los resultados de un análisis factorial que implica a esta prueba. Luego, si se conoce la confiabilidad de la prueba, su especificidad puede encontrarse por sustracción. Un análisis factorial ilustrativo debe aclarar estas cuestiones.

Una forma de comenzar un análisis factorial de las calificaciones de *n* personas en un grupo de pruebas es calcular las correlaciones entre todas las pruebas y acomodarlas en la forma de

TABLA A.3 Matriz de correlaciones promedio entre las subpruebas de la WISC-III

SUBPRUEBA	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Información		.66	.57	.70	.56	.34	.47	.21	.40	.48	.41	.35	.18
2. Semejanzas	.66		.55	.69	.59	.34	.45	.20	.39	.49	.42	.35	.18
3. Aritmética	.57	.55		.54	.47	.43	.39	.27	.35	.52	.39	.41	.22
4. Vocabulario	.70	.69	.54		.64	.35	.45	.26	.40	.46	.41	.35	.17
5. Comprensión	.56	.59	.47	.64		.29	.38	.25	.35	.40	.34	.34	.17
6. Retención de dígitos	.34	.34	.43	.35	.29		.25	.23	.20	.32	.26	.28	.14
7. Figuras incompletas	.47	.45	.39	.45	.38	.25		.18	.37	.52	.49	.33	.24
8. Codificación	.21	.20	.27	.26	.25	.23	.18		.28	.27	.24	.53	.15
9. Ordenamiento de figuras	.40	.39	.35	.40	.35	.20	.37	.28		.41	.37	.36	.23
10. Diseño con cubos	.48	.49	.52	.46	.40	.32	.52	.27	.41		.61	.45	.31
11. Ensamble de objetos	.41	.42	.39	.41	.34	.26	.49	.24	.37	.61		.38	.29
12. Búsqueda de símbolos	.35	.35	.41	.35	.34	.28	.33	.53	.36	.45	.38		.24
13. Laberintos	.18	.18	.22	.17	.17	.14	.24	.15	.23	.31	.29	.24	

una matriz. Esto se ha hecho en la tabla A.3 con las correlaciones promedio entre las subpruebas de la WISC III para todas las edades en la muestra de estandarización de la prueba (n = 1,880). Advierta que la matriz es simétrica; es decir, las correlaciones en una hilera determinada son idénticas a las de la columna correspondiente. Además, no hay entradas en la diagonal que va de la esquina superior izquierda a la esquina inferior derecha de la matriz.

La decisión concerniente a qué valores colocar en la diagonal de la matriz —las confiabilidades de las pruebas, las estimaciones de sus comunalidades o todos los 1.00—depende del procedimiento particular de análisis factorial o de la teoría que siga el investigador. En un tipo de procedimiento de factorización, el *método central*, en la diagonal de la matriz de correlación se colocan las estimaciones de las comunalidades de las pruebas. Por otro lado, el *método de componentes principales* requiere colocar los 1.00 en la diagonal. Sin extenderse en la cuestión de qué entradas en la diagonal son mejores, debe enfatizarse que la elección afecta tanto el número de factores extraídos como los pesos obtenidos (*cargas factoriales*) de cada prueba en cada factor. El siguiente análisis factorial fue realizado siguiendo el método de componentes principales, usando 1.00 en la diagonal de la matriz de correlación.

Factorización de la matriz de correlación. El resultado inmediato de un análisis factorial típico es una matriz de factores original (sin rotar) como la que se presenta en las columnas *A*, *B* y *C* de la tabla A.4. Observe que el análisis factorial ha reducido el número de variables o dimensiones psicológicas de 13, que es el número total de subpruebas en la WISC-III, a tres, el número de factores comunes extraído. Los números decimales en cada columna de la matriz de factores son las cargas de las 13 subpruebas de la WISC-III en ese factor. Por ejemplo, la subprueba de Información tiene una carga de .78 en el factor *A*, pero tiene cargas de sólo −.33 y .03 en los factores *B* y *C*. Cada carga de factor es la correlación entre una subprueba particular y ese factor. El

TABLA A.4 Matrices de factores originada y rotada

	MARIZ DE FACTORES ORIGINAL			MARIZ DE FACTORES ROTADA			
Subprueba	*A*	*B*	*C*	*A′*	*B′*	*C′*	*Comunalidad*
Información	.78	–.33	.03	.80	.25	.09	.71
Semejanzas	.78	–.34	.02	.81	.25	.08	.72
Aritmética	.74	–.10	.12	.65	.26	.28	.57
Vocabulario	.79	–.34	.10	.83	.19	.13	.74
Comprensión	.70	–.29	.14	.75	.14	.16	.61
Retención de dígitos	.51	–.02	.28	.45	.06	.36	.34
Figuras incompletas	.66	.01	–.35	.43	.61	.02	.56
Codificación	.44	.55	.54	.10	.09	.88	.79
Ordenamiento de figuras	.60	.16	–.08	.34	.45	.27	.39
Diseño con cubos	.75	.18	–.26	.41	.66	.22	.65
Ensamble de objetos	.66	.22	–.36	.31	.71	.14	.62
Búsqueda de símbolos	.62	.48	.30	.23	.32	.74	.70
Laberintos	.37	.42	–.45	–.06	.71	.11	.52

cuadrado de la carga de una subprueba determinada en un factor es la proporción de la varianza total de las calificaciones de la subprueba que puede ser explicada por ese factor. De este modo, $(.78)^2 = .61$ significa que 61% de la varianza de las calificaciones en la subprueba de Información puede ser explicada por el factor *A*. Sólo $(-.33)^2 = .11$, u 11%, de la varianza de las calificaciones de la subprueba de Información puede ser explicada por el factor *B*, y $(.03)^2 = .0009$, o .09%, de la varianza de la subprueba de Información puede ser explicada por el factor *C*.

La suma de los productos cruzados de las cargas factoriales correspondientes de dos subpruebas cualesquiera en la tabla A.4 es una estimación de la correlación entre esas dos subpruebas. Por ejemplo, se estima que la correlación entre las subpruebas de Información y Aritmética a partir de las cargas en la matriz de factores original es $.78(.74) + (-.33)(-.10) + (.03)(.12) = .61$. Ésta es una aproximación bastante cercana a la correlación real de .57 (vea la tabla A.3). La precisión con la que se reproduce la matriz de correlación mediante estimaciones determinadas a partir de las cargas factoriales depende del grado en que los factores obtenidos expliquen la varianza total entre las subpruebas.

Rotación de los factores. Es posible aplicar un proceso conocido como *rotación de factores* a la matriz de factores original a fin de incrementar el número de cargas positivas altas y bajas en las columnas de la matriz de factores. El resultado es una configuración más simple de las cargas factoriales, lo que facilita la interpretación de los factores. Dependiendo del método particular de rotación seleccionado, pueden obtenerse factores no correlacionados (*ortogonales*) o correlacionados (*oblicuos*). Algunos analistas factoriales prefieren la rotación ortogonal mientras que a otros les agrada la rotación oblicua. La matriz de factores rotados mostrada en la tabla A.4 (columnas A', B' y C') se produjo mediante una rotación ortogonal (varimax) de la matriz de factores original, por ello los factores rotados son no correlacionados.

Interpretación de los factores rotados. Una vez completados los cálculos estadísticos implicados en la factorización de la matriz de correlación y la rotación de los factores extraídos, estamos listos para examinar el patrón de cargas altas y bajas de cada prueba en cada factor. Entre más alta sea una carga particular, más importante es el factor en la prueba dada. Como se muestra en la tabla A.4, las subpruebas de Información, Semejanzas, Vocabulario y Comprensión tienen cargas de más de .70 en el factor A'. Debido a que éstas son subpruebas verbales, el factor A' puede denominarse factor *verbal*. Otras subpruebas también tienen cargas apreciables en el factor A', así que este factor se acerca a lo que se entiende por un *factor cognoscitivo general* (*g*). Las subpruebas de Figuras incompletas, Diseño con cubos y Ensamble de objetos tienen cargas de moderadas a altas en el factor B'. Considerando los tipos de tareas que comprenden esas tres subpruebas, el factor B' puede denominarse factor *espacial-perceptual* o de *imágenes espaciales*. Por último, las subpruebas de Codificación y Búsqueda de símbolos, ambas implican transformar un conjunto de símbolos abstractos a otro, tienen cargas altas en el factor C'. Éste parece ser un factor bastante específico que comprende rapidez perceptual, precisión y libertad de pasatiempos.

Comunalidad y especificidad. La última columna de la tabla A.4 contiene las comunalidades de las 13 subpruebas de la WISC-III, calculadas como la suma de los cuadrados de las cargas

factoriales rotadas en una subprueba determinada. Por ejemplo, la comunalidad de la subprueba de Información es $(.80)^2 + (.25)^2 + (.09)^2 = .71$, por lo que 71% de la varianza de las calificaciones de la subprueba de Información puede ser explicada por los factores A', B' y C'. Si se conoce la confiabilidad de la subprueba de Información, puede usarse la fórmula A.8 para calcular su especificidad. Además, al restar la comunalidad de 1.00 se obtiene la proporción de la varianza total de la subprueba que puede atribuirse a una combinación de factores específicos y la varianza de error. Para la subprueba de Información esta cifra es $1.00 - .71 = .29$; es decir, 29% de la varianza total de las calificaciones en la subprueba de Información puede ser explicada por factores específicos y errores de medición. Sabiendo que la confiabilidad estimada de la subprueba de Información es .84, podemos restar su comunalidad (.71) y encontrar que su especificidad es .13 (vea la fórmula A.8).

RESUMEN

El análisis estadístico de las calificaciones de las pruebas comienza con la elaboración de una distribución de frecuencia del número de personas que obtienen cada calificación o cuyas calificaciones caen dentro de un intervalo especificado. Las distribuciones de frecuencia pueden representarse de manera gráfica como histogramas o polígonos de frecuencia. La curva normal es un polígono de frecuencia teórico que resulta básico para gran parte de la teoría de las pruebas y se utiliza con diversos propósitos. Las distribuciones de frecuencia asimétricas, no normales, pueden tener un sesgo a la derecha (con sesgo positivo) o a la izquierda (con sesgo negativo).

Tres medidas de la tendencia central o promedio de un grupo de calificaciones, la moda, la mediana y la media, pueden calcularse a partir de las calificaciones crudas o de una distribución de frecuencia. La moda es la calificación que ocurre con mayor frecuencia, la mediana es el valor por debajo del cual cae 50% de las calificaciones y la media aritmética es la suma de las calificaciones dividida entre el número de calificaciones. Tres medidas de la variabilidad o dispersión de un grupo de calificaciones son el rango, el rango semiintercuartilar y la desviación estándar. De éstas, la desviación estándar es la medida de variabilidad más popular y más apropiada cuando la media aritmética es el promedio reportado. Para propósitos de comparación e interpretación, las calificaciones crudas pueden convertirse a calificaciones estándar *z*, percentiles y otras calificaciones transformadas.

El coeficiente de correlación producto-momento, el cual es un número entre -1.00 (correlación negativa perfecta) y $+1.00$ (correlación positiva perfecta), es una medida de la magnitud y dirección de la relación entre dos variables. Una correlación significativa entre dos variables facilita la predicción de la calificación de una persona en una variable a partir de su calificación en la otra variable. Sin embargo, no debe suponerse que una correlación elevada entre dos variables implica una conexión causal entre ellas. Aunque la causalidad implica correlación, la correlación no implica causalidad.

Las correlaciones entre variables pueden usarse en análisis de regresión lineal simple y múltiple para emitir pronósticos de las calificaciones en una variable dependiente (*Y* o criterio) a partir de las calificaciones en una o más variables independientes (*X* o pronosticadoras). Los procedimientos de correlación también se usan en el análisis factorial para determinar las dimensiones o factores que diferentes pruebas tienen en común. El análisis factorial de las califi-

caciones obtenidas por una muestra grande de personas en un grupo de pruebas o reactivos consiste en extraer los factores, rotar los ejes de los factores, e interpretar los factores resultantes. Los factores son interpretados inspeccionando las cargas de las diversas pruebas en el factor. El cálculo de la comunalidad (varianza de factores comunes) y la especificidad (varianza de factores específicos) también puede contribuir al proceso de interpretación de factores.

PREGUNTAS Y ACTIVIDADES

1. Considere la siguiente distribución de frecuencia de las calificaciones obtenidas por un grupo de 50 estudiantes en una prueba:

INTERVALO DE LA CALIFICACIÓN DE LA PRUEBA	NÚMERO DE ESTUDIANTES
96–100	6
91–95	8
86–90	15
81–85	10
76–80	7
71–75	4

Trace un histograma y un polígono de frecuencia sobreimpuesto de esta distribución de frecuencia. Calcule luego la media aritmética, la mediana, la desviación estándar, el percentil 25°, el percentil 75° y el rango semiintercuartilar de las calificaciones.

2. Usando la tabla del apéndice B, encuentre el porcentaje del área bajo la curva normal que cae por debajo de cada una de las siguientes calificaciones z: -2.575, -2.33, -1.96, -1.645, .00, 1.645, 1.96, 2.33 y 2.575. A continuación encuentre las calificaciones z debajo de las cuales cae 10, 20, 30, 40, 50, 60, 70, 80 y 90 por ciento del área bajo la curva normal.

3. Considere los siguientes pares de calificaciones X, Y de 30 personas:

X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
32	46	28	23	37	28	36	21	42	27
35	26	32	20	27	13	31	14	39	46
20	8	45	24	37	22	35	18	34	16
41	42	29	13	23	34	43	47	33	30
25	28	46	40	30	31	34	27	29	26
38	25	40	37	36	39	39	32	24	7

Calcule los siguientes estadísticos: medias aritméticas y desviaciones estándar de X y Y, correlación producto-momento entre X y Y, y línea de regresión para predecir Y a partir de X. Represente gráficamente (*diagrama de dispersión*) los puntos X, Y, y dibuje la línea de regresión de Y en X.

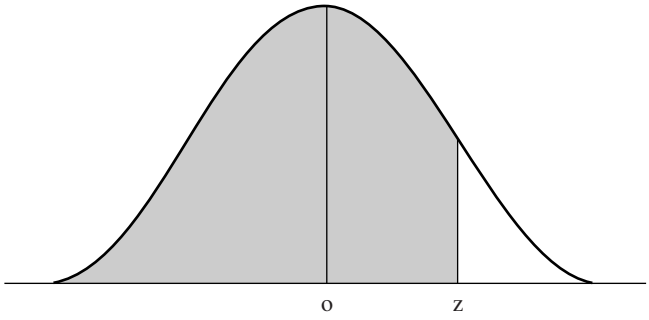
4. Cada vez que la distribución de frecuencia de un grupo de calificaciones muestra un sesgo marcado en una dirección positiva (a la derecha) o negativa (a la izquierda), se considera que la mediana es una medida mejor y menos sesgada de la tendencia central que la media aritmética. ¿Por qué?

5. ¿Cuál es el propósito de conducir un análisis de regresión múltiple en un conjunto de datos psicométricos? ¿Cuál es el propósito de realizar un análisis factorial? Consulte las bases de datos PsycLIT o PsycINFO de los últimos años y encuentre dos estudios en los cuales se haya empleado un procedimiento de regresión múltiple y otros dos estudios donde se haya usado un análisis factorial. Resuma el procedimiento usado y los resultados obtenidos.

APÉNDICE B

ÁREAS BAJO LA CURVA NORMAL

Para encontrar la proporción del área bajo la curva normal que cae por debajo de un valor específico de *z*, localice el valor de *z* en la primera columna y la fila superior de la tabla. El número decimal en la intersección de la fila y la columna apropiadas es la proporción del área bajo la curva. Por ejemplo, para encontrar el área bajo $z = 1.57$, busque la intersección de 1.5 en la primera columna y .07 en la fila superior. El valor resultante es .9418, de modo que 94.18% del área bajo la curva cae por debajo de $z = 1.57$. A la inversa, para localizar el valor de *z* por debajo del cual cae una proporción específica del área bajo la curva, empiece por buscar dicha proporción en el cuerpo de la tabla. Entonces busque el valor de *z* en la fila y columna correspondientes. Por ejemplo, para encontrar el valor de z por debajo del cual cae 67% del área bajo la curva, se empieza por localizar .6700 en el cuerpo de la tabla. Está en la intersección de la fila .4 y la columna .04, de modo que el valor *z* correspondiente es .44.



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
–3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00117	.00104	.00100
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
–1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
–1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
–1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
–1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
–1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
–1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
–1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
–1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
–1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
–0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
–0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
–0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
–0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
–0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
–0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
–0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
–0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
–0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
–0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

APÉNDICE C

DISTRIBUIDORES COMERCIALES DE MATERIAL DE EVALUACIÓN PSICOLÓGICA Y EDUCATIVA

Academic Therapy Publications, 20 Commercial Boulevard, Novatao, CA 93939-6191. Teléfonos: 415-883-3314, 800-422-7249. FAX: 415-883-3720. Sitio Web: www.atpub.com.

Allyn & Bacon, Department 893, 160 Gould Street, Needham Heights, MA 02194-2310.

The American College Testing Program (ACT), P.O. Box 168, Iowa City, IA 52243-0168. Teléfono: 319-337-1000. FAX: 319-337-1551. Sitio Web: www.act.org.

American Guidance Service, Inc. (AGS), 4201 Woodland Road, Circle Pines, MN 55014-1796. Teléfonos: 612-786-4343, 800-328-2560. FAX: 612-786-9077. E-mail: agsmail@agsnet. Sitio Web: www.agsnet.com.

American Psychiatric Press, Inc., 1400 K Street NW, Suite 1101, Washington, DC 20005.

Assessment Systems Corporation, 2233 University Avenue, Suite 200, St. Paul, MN 55114-1629. Teléfono: 651-647-9220. FAX: 651-647-0412. E-mail: info@assess.com. Sitio Web: www.assess.com.

Australian Council for Educational Research, Ltd., 347 Camberwell Road (Private Bag 55), Camberwell, Victoria, Australia 3124. Teléfono: (03) 9835 7447. FAX: (03) 9835 7499. E-mail: sales@acer.edu.au. Sitio Web: www.acer.edu.au.

Behavior Science Systems, Inc., P.O. Box 580274, Minneapolis, MN 55458. Teléfono: 612-929-6220. FAX: 612-920-4925.

Dr. Martin M. Bruce, 22516 Caravelle Circle, Boca Raton, FL 33433. Teléfono: 561-393-2428. FAX: 561-362-6185. E-mail: brucepubl@aol.com.

Center for Applied Linguistics, 1118 22nd Street NW, Washington, DC 20037.

Center for the Study of Attitudes Toward Persons with Disabilities, Hofstra University; Hofstra University, Hempstead, NY 11549-1000; Teléfono: 516-463-6600.

The College Board, 45 Columbus Avenue, New York, NY 10023-6992. Teléfono: 212-713-8390. Sitio Web: www.collegeboard.org.

Consulting Psychologists Press, Inc. (CPP), 3803 East Bayshore Road, P.O. Box 10096, Palo Alto, CA 94303. Teléfonos: 650-969-8901, 800-624-1765. FAX: 650-969-8608. Sitio Web: www.cpp-db.com.

Consulting Resource Group International, Inc., # 886, 200 West Third Street, Sumas, WA 98295-8000.

C.P.S., Inc., P.O. Box 83, Larchmont, NY 10538. Teléfono: 914-833-1633. FAX: 914-833-1633.

CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940-5703. Teléfonos: 800-538-9547, 899-682-922 (en CA). FAX: 800-282-0266. TDD: 800-217-9190. Sitio Web: www.ctb.com.

Denver Developmental Materials, Inc., P.O. Box 6919, Denver, CO 80206-0919. Teléfono: 800-419-4729. FAX: 303-355-5622.

DLM Resources, One DLM Park, Allen, TX 75002. Teléfono: 800-527-4747.

Department of Defense, Manpower Data Center, DoD Center Monterey Bay, 400 Gigling Road, Seaside, CA 93955-6771. Teléfono: 831-583-2400, ext. 4284.

Department of Research Assessment and Training, 1051 Riverside Drive, Unit 123, New York, NY 10032. Teléfono: 212-543-5536. FAX: 212-543-5386.

EdITS/Educational and Industrial Testing Service, P.O. Box 7234, San Diego, CA 92167-0234. Teléfonos: 619-222-1666, 800-416-1666. FAX: 619-226-1666. E-mail: edits@worldnet.att.net. Sitio Web: www.edits.net.

Educational Records Bureau, 220 East 42nd Street, Suite 100, New York, NY 10017. Teléfono: 212-672-9800. Sitio Web: www.erbtest.org.

Educational Testing Service, Rosedale Road, Princeton, NY 08541-0001. Teléfono: 609-921-9000. FAX: 609-734-5410. Sitio Web: www.ets.org.

Educators Publishing Service, Inc., 31 Smith Place, Cambridge, MA 02138-1089. Teléfono: 800-225-5750. FAX: 617-547-0412. E-mail: cps@epsbooks.com. Sitio Web: www.epsbooks.com.

Elbern Publications, P.O. Box 98497, Columbus, OH 43209. Teléfono: 614-235-2643. FAX: 614-237-2637.

ETS Test Collection, Educational Testing Service, Princeton, NJ 08541. Teléfono: 609-734-5686. Sitio Web: www.ets.org.

Family Social Science, 290 McNeal Hall, University of Minnesota, 1985 Buford Avenue, Saint Paul, MN 55108. Teléfono: 612-625-7250. FAX: 612-625-4227.

General Educational Development Testing Service of the American Council on Education (GED), One Dupont Circle NW, Suite 250, Washington, DC 20036-1163. Teléfono: 202-939-9490. FAX: 202-775-8578. E-mail: ged@ace.nche.edu. Sitio Web: www.gedtest.org.

GIA Publications, Inc., 7404 S. Mason Avenue, Chicago, IL 60638. Teléfono: 708-496-3800. FAX: 708-496-3828.

Harcourt Brace Educational Measurement, 555 Academic Court, San Antonio, TX 78204-2498. Teléfono: 512-299-1061, 800-228-0752. FAX: 800-232-1223. E-mail: customer_service@hbtpc.com. Sitio Web:www.hbem.com.

Harvard University Press, 79 Garden Street, Cambridge, MA 02138. Teléfono: 800-448-4083. E-mail: hup@harvard.edu. Sitio Web: www.hup.harvard.edu.

Hawthorne Educational Services, Inc., 800 Gray Oak Drive, Columbia, MO 65201. Teléfono: 800-542-1673. FAX: 800-442-9509.

Hilson Research, Inc., P.O. Box 150239, 82-88 Abingdon Road, Kew Gardens, NY 11415-0239. Teléfono: 800-926-2258. Fax: 718-849-6238.

Hodder & Stoughton Educational, 338 Euston Road, London, NWl 3BH, England. Teléfono: +44 w0 7873 6286. FAX: +44 20 7873 6299. E-mail: lucy.johnson@hodder.co.uk.

Hogan Assessment Systems, Inc., P.O. Box 521176, Tulsa, OK 74152-1176. Teléfono: 918-749-0632. FAX: 918-749-0635. E-mail: aferg@webzone.net. Sitio Web: www.hoganassessments.com.

Hogrefe & Huber Publishers, P.O. Box 2487, Kirkland, WA 98083. Teléfono: 800-228-3749. FAX: 424-823-8324. E-mail: hh@hhpub.com. Sitio Web: www.hhpub.com.

IDS Publishing P.O. Box 389, Worthington, OH 43085.

Industrial Psychology International, Ltd., 4106 Firestone Road, Champaign, IL 61821. Teléfono: 800-747-1229. FAX: 217-398-5798.

Institute for Personality and Ability Testing (IPAT), P.O. Box 1188, Champaign, IL 61824-1188. Teléfono: 800-225-IPAT. E-mail: custserv@ipat.cfom. Sitio Web: www.ipat.com.

JIST Works, Inc., 720 North Park Avenue, Indianapolis, IN 46202-3431. Teléfono: 800-648-5478. FAX: 800-547-8329. E-mail: jistworks@aol.com. Sitio Web: www.jist.com.

LinguiSystems, Inc., 3100 4th Avenue, East Moline, IL 61244-9700. Teléfonos: 800-776-4332, 309-755-2300. FAX: 309-755-2377. TDD: 800-933-8331. E-mail: service@linguisystem.com.

Management Research Institute, Inc., 11304 Spur Wheel Lane, Potomac, MD 20854. Teléfono: 301-299-9200. FAX: 301-299-9227. E-mail: mrieaf@aol.com.

Marathon Consulting and Press, 797 South Ashburton Road, Columbus, OH 43227-1027.

McCarron-Dial Systems, P.O. Box 45628, Dallas, TX 75245. Teléfono: 214-634-2863. FAX: 214-634-9970. E-mail: mds@mccarrondial.com. Sitio Web: mccarrondial.com.

MetriTech, Inc., 4106 Fieldstone Road, P.O. Box 6479, Champaign, IL 61826-6479. Teléfono: 217-398-4868. FAX: 217-398-5798. E-mail: ipi@metritech.com. Sitio Web: www.metritech.com.

Mind Garden, Inc., 1690 Woodside Road, Suite 202, Redwood City, CA 94061. Teléfono: 650-261-3500. FAX: 650-261-3505. E-mail: info@mindgarden.com.

Multi-Health Systems, Inc., 908 Niagara Falls Boulevard, North Tonawanda, NY 14120-2060. Teléfono: 800-456-3003 o 416-492-2627. FAX: 888-540-4484 o 416-492-3343. E-mail: customerservice@mhs.com. Sitio Web: www.mhs.com.

National Career Assessment Services, Inc., 601 Visions Parkway, P.O. Box 277, Adel, IA 50003. Teléfono: 800-314-8972. FAX: 515-993-5422. E-mail: ncasi@ncasi.com. Sitio Web: www.kuder.com.

NCS London House, 9701 West Higgins Road, Suite 170, Rosemont, IL 60018-4720. Teléfono: 800-221-8378. Sitio Web: londonhouse.ncspearson.com.

NCS Assessments, P.O. Box 1416, Minneapolis, MN 55440. Teléfono: 800-627-7271. FAX: 800-632-99011. Sitio Web: assessments.ncspearson.com.

NFER-Nelson Publishing Company, Ltd., Darville House, 2 Oxford Road East, Windsor-Berkshire, SL4 lDF, England.

Helen Orvaschel, Nova University Center for Psychological Studies, 3301 College Avenue, Fort Lauderdale, FL 33314.

Oxford Psychologists Press, Ltd., Lambourne House 311-321 Banbury Road, Oxford OX27JH, United Kingdom. Teléfono: +44 1865 404500. FAX: +44 1865 310365. Sitio Web: 333.opp.co.uk.

Personnel Decisions International, 2000 Plaza VII Tower, 45 South Seventh Street, Minneapolis, MN 55402-1608.

Personnel Press, 191 Spring Street, Lexington, MA 02173.

Pfeiffer & Company International Publishers, 2780 Circleport Drive, Erlander, KY 41018. Teléfonos: 800-274-4434, 606-647-3030. FAX: 800-569-0443.

pro.ed, 8700 Shoal Creek Blvd., Austin, TX 78757-6897. Teléfonos: 512-451-3246, 800-897-3202. FAX: 800-397-7633. Sitio Web: www.proedinc.com.

Psychological Assessment Resources, Inc. (PAR), P.O. Box 998, Odessa, FL 33556-0998. Teléfonos: 813-968-3003, 800-331-8378. FAX: 800-727-9329. Web URL: www.parinc.com.

The Psychological Corporation, 555 Academic Court, San Antonio, TX 78204-2498. Teléfono: 800-211-8378. FAX: 800-232-1223. E-mail: customer_service@HBTPC.com. Sitio Web: www.PsychCorp.com.

Psychological Publications, Inc., P.O. Box 3577, Thousand Oaks, CA 91361-0577. Teléfono: 800-345-8378. FAX: 805-373-1753. E-mail: TJTA@aol.com. Sitio Web: www.TJTA-.com.

Psychological Services, Inc., 100 West Broadway, Suite 1100, Glendale, CA 91210. Teléfono: 818-244-0033. FAX: 818-247-7223. E-mail: testinfo@psionline.com. Sitio Web: www.psionline.com.

Psychological Test Specialists, Box 9229, Missoula, MT 59807.

Psychologists and Educators, Inc., Sales Division, P.O. Box 513, Chesterfield, MO 63006. Teléfono: 314-536-2366. FAX: 314-434-2331.

Psychometric Affiliates, P.O. Box 807, Murfreesboro, TN 37133-0807. Teléfono: 615-898-8265. FAX: 615-890-6296. E-mail: jheritage@al.mtsu.edu.

Publishers Test Service, 2500 Garden Road, Monterey, CA 93940-5379. Teléfono: 800-538-9547.

Purdue Research Foundation, ATTN: William K. Lebold, Educational Research & Info System, Engineering and Administration Bldg., Purdue University, West Lafayette, IN 47907.

Reitan Neuropsychology Laboratory/Press, P.O. Box 66080, Tucson, AZ 85728-6080. Teléfono: 520-882-2022. FAX: 520-884-0040. E-mail: reitanlab@aol.com.

Research Press, Dept. G., P.O. Box 9177, Champaign, IL 61826.

Riverside Publishing, 425 Spring Lake Drive, Itasca, IL 60143-2079. Teléfono: 800-323-9540. FAX: 312-693-0325. Sitio Web: www.riverpub.com.

Scholastic Testing Service, Inc. (STS), 480 Meyer Road, P.O. Box 1056, Bensenville, IL 60106-1617. Teléfono: 800-766-7150 u 800-642-6STS. FAX: 630-766-8054. E-mail: ststestin-g@email.com. Sitio Web: www.ststeesting.com.

Sigma Assessment Systems, Inc., 511 Fort Street, Suite 435, P.O. Box 610984, Port Huron, MI 48061-0984. Teléfono: 800-265-1285. FAX: 800-361-9411. E-mail: SIGMA@sigmaas-sessment systems.com. Sitio Web: www.sigmaassessmentsystems.com.

Slosson Educational Publications, Inc., P.O. Box 280, East Aurora, NY 14052-0280. Teléfonos: 716-652-0930, 800-828-4800. FAX: 800-655-3840. Sitio Web: www.slosson.com.

SOI Systems, P.O. Box D, 45755 Goodpasture Road, Vida, OR 97488. Teléfono: 541-896-3936. FAX: 541-896-3983. E-mail: rmeeker@soisystems.com. Sitio Web: www.soisystems.com.

Special Child Publications, P.O. Box 33548, Seattle, WA 98133.

Stoelting Co., Oakwood Center, 620 Wheat Lane, Wood Dale, IL 60191. Teléfono: 630-860-9700. FAX: 630-860-9775. E-mail: psychtests@stoeltingco.com. Sitio Web: www.stoel-tingco.com.

Swets Test Services, P.O. Box 820, 2160 Sz Lisse, The Netherlands. Teléfono: +31 252 435375. FAX: +31 252 435671. E-mail: dvants@swets.nl. Sitio Web: www.swetstest.nl.

Talico, Inc., 2320 S. Third Street, Suite 5, Jacksonville, FL 32250-4057.

21st Century Assessment, P.O. Box 608, South Pasadena, CA 91031. Teléfono: 800-374-2100. FAX: 626-441-0614.

University of Minnesota Press, Test Division, Mill Place, Suite 290, 111 Third Avenue, South, Minneapolis, MN 55401-2520.

University of Vermont Department of Psychiatry, One South Prospect Street, Burlington, VT 05401-3456. Teléfono: 802-656-8313. FAX: 802-656-2602.

U. S. Employment Service, Division of Program Planning and Operations, U. S. Department of Labor, 601 D Street, NW, Washington, DC 20213.

U. S. Military Entrance Processing Command Testing Directorate, 2500 Green Bay Road, North Chicago, IL 60064-3094.

Vocational Research Institute (VRI), 1528 Walnut Street, Suite 1502, Philadelphia, PA 19102. Teléfono: 800-874-5387. FAX: 215-875-0198. E-Mail: info@vri.org. Sitio Web: www.-vri.org.

Western Psychological Services (WPS), 12031 Wilshire Boulevard, Los Ángeles, CA 90025-1251. Teléfono: 310-478-2061, 800-648-9957. FAX: 310-478-7838.

Wide Range, Inc., P.O. Box 3410, Wilmington, DE 19804-0250. Teléfonos: 302-652-4990, 800-221-9728. FAX: 302-652-1644. Sitio Web: www.widerange.com.

Wonderlic Personnel Test, Inc., 1795 N. Butterfield Road, Libertyville, IL 60048-11238. Teléfono: 800-323-3742. FAX: 847-680-9492. Sitio Web: www.wonderlic.com.

APÉNDICE D

SITIOS WEB DE ORGANIZACIONES INTERESADAS EN LA EXAMINACIÓN PSICOLÓGICA Y LA EVALUACIÓN

American Counseling Association (ACA): www.counseling.org

American Council on Education (ACE): www.acenet.edu

American Educational Research Association: www.aera.net

American Psychological Association: www.apa.org/science/testing.html

American Speech–Language–Hearing Association (ASHA): www.asha.org

The Association for Assessment in Counseling: aac.ncat.edu

Association of Test Publishers: www.testpublishers.org

Buros Institute of Mental Measurements: www.unl.edu/buros

College PowerPrep: www.powerprep.com

Educational Resources Information Center (ERIC): www.ericae.net\

Educational Testing Service (ETS): www.ets.org

ERIC Clearinghouse on Assessment and Evaluation: www.ericae.net

Graduate Record Examinations (GRE): www.gre.org

International Personnel Management Association (IPMAAC): www.ipmaac.org

Kaplan Inc.: www.kaplan.com

National Academy of Science's Board on Testing and Assessment (BOTA): www4.nationalacademies.org/dbasse/bota.nsf

National Association of School Psychologists (NASP): www.naspweb.org

National Association of Test Directors (NATD): www.naspweb.org

National Center for Research on Evaluation, Standards, and Student Testing (CRESST): www.cse.ucla.edu/

National Council on Measurement in Education (NCME): www.ncme.ed.uiuc.edu

Princeton Review: www.review.com

Society for Industrial and Organizational Psychology (SIOP): www.siop.org

Test.com, Inc.: www.test.com

GLOSARIO

Acomodación. (1) En la teoría de J. Piaget del desarrollo cognoscitivo, la modificación de los esquemas como resultado de la experiencia. (2) Un cambio en la forma en que se administra una prueba o en la forma en que se permite responder al examinado.

Actitud. Tendencia a reaccionar de manera positiva o negativa a algún objeto, persona o situación.

Afasia. Defecto en la habilidad para comunicarse (mediante el habla, por escrito o por signos) y/o para comprender el lenguaje hablado o escrito, ocasionado por enfermedad o lesión del cerebro.

Agnosia. Incapacidad parcial o total para reconocer estímulos sensoriales.

Ajuste. Capacidad para enfrentar las situaciones sociales y obtener la satisfacción de las necesidades.

Alexia. Deterioro de la habilidad para leer.

Análisis de contenido. Método para estudiar y analizar comunicaciones escritas (u orales) de manera sistemática, objetiva y cuantitativa para evaluar ciertas variables psicológicas.

Análisis de reactivos. Término general para los procedimientos diseñados para evaluar la utilidad o validez de un grupo de reactivos de una prueba.

Análisis de regresión lineal. Procedimiento para determinar la ecuación algebraica de la línea de mejor ajuste para predecir las calificaciones en una variable dependiente a partir de una o más variables independientes.

Análisis de regresión múltiple. Método estadístico para analizar las contribuciones de dos o más variables independientes a la predicción de una variable dependiente.

Análisis de trabajo. Término general para procedimientos usados para determinar los factores o tareas que componen un trabajo. El análisis de trabajo por lo regular se considera un requisito previo a la elaboración de una prueba para predecir el desempeño en el trabajo.

Análisis de varianza múltiple abstracta (MAVA). Procedimiento estadístico, desarrollado por R. B. Cattell, para determinar los efectos relativos de la herencia y el ambiente en una característica particular de personalidad.

Análisis del comportamiento. Procedimientos que se centran en la descripción objetiva de un comportamiento particular y en la identificación de los antecedentes y las consecuencias de esta conducta. El análisis del comportamiento puede realizarse con propósitos de investigación o para obtener información al planear un programa de modificación del comportamiento.

Análisis factorial. Procedimiento matemático para analizar una matriz de correlaciones entre mediciones con el fin de determinar qué factores (constructos) son suficientes para explicar las correlaciones.

Anchura de banda. Término de L. J. Cronbach para referirse al rango de criterios predecibles a partir de una prueba; entre mayor sea el número de criterios que una prueba puede predecir, mayor es la anchura de banda. (Vea *fidelidad.*)

Ansiedad por una prueba. Ansiedad en una situación de prueba.

Apareamiento clasificado. Apareamiento no aleatorio entre individuos que poseen características similares.

Apraxia. Incapacidad para hacer movimientos voluntarios (por ejemplo, incapacidad para usar un objeto de manera apropiada), aunque no hay parálisis u otro deterioro en las habilidades sensoriales o motrices.

Aproximación ABC. Aproximación de evaluación conductual que implica la identificación de los eventos antecedentes (A) y de las consecuencias (C) de la conducta (B). La conducta es modificada controlando A y cambiando C.

Aproximación clínica (impresionista). Enfoque de diagnóstico y predicción conductual en el cual los psicólogos o psiquiatras asignan sus propios pesos sentenciosos a las variables predictoras y luego los combinan de manera subjetiva para hacer diagnósticos y pronósticos.

Aptitud académica. La habilidad para aprender tareas de tipo escolar; también se conoce como *aptitud escolar*. Muchas pruebas de inteligencia son básicamente medidas de aptitud académica.

Aptitud. Capacidad para aprender a realizar una tarea o habilidad particular. Tradicionalmente, se pensaba que la aptitud dependía más del potencial innato que de la experiencia y la práctica.

Área de Broca. Área en la parte frontal izquierda de la corteza cerebral, relacionada con el control del habla. Los pacientes con daño en el área de Broca tienen problemas para enunciar las palabras correctamente y hablan de manera lenta y dificultosa. (Vea *Área de Wernicke.*)

Área de Wernicke. Área en el hemisferio cerebral izquierdo relacionado con la comprensión del lenguaje.

Los pacientes con daño en el área de Wernicke pueden oír las palabras pero no logran entender su significado. (Vea *área de Broca.*)

Asimilación. En la teoría de J. Piaget del desarrollo cognoscitivo, el proceso de ajustar las experiencias nuevas a las estructuras mentales (*esquemas*) ya existentes.

Ataxia. Pérdida de coordinación muscular, en particular de las extremidades.

Audiómetro. Instrumento para medir la agudeza auditiva que presenta tonos puros de diversas intensidades y frecuencias en el rango normal de audición. La audición se prueba en cada oído. Los resultados se trazan en forma de un audiograma, una gráfica de la agudeza auditiva del examinado en cada frecuencia y para cada oído.

Autoconcepto. Evaluación que hace una persona de su habilidad para realizar con éxito una tarea particular en cierta situación.

Autoeficacia. Juicio que hace una persona concerniente a su habilidad para realizar con éxito una tarea particular en cierta situación.

Banda percentilar. Un rango de rangos percentilares dentro del cual hay una probabilidad especificada de que caerá la verdadera calificación de un individuo en una prueba.

Batería de aptitudes múltiples. Batería de pruebas conormadas diseñadas para evaluar las capacidades mentales.

Batería de tests. Grupo de tests de aptitudes o aprovechamiento que miden diferentes cosas, pero que se estandarizaron en la misma muestra, permitiendo hacer comparaciones del desempeño de una persona en áreas diferentes.

Calificación analítica. Procedimiento de calificación para pruebas de ensayo en el cual se asignan diferentes puntuaciones a aspectos de contenido y estilo de las respuestas del examinado.

Calificación Apgar. Calificación determinada al minuto y a los cinco minutos del nacimiento para evaluar a los neonatos. Se asigna una calificación de 0 a 2 a mediciones del ritmo cardiaco, respiración, tono muscular, reflejos y color. Una suma de calificaciones entre 7 y 10 es normal para los recién nacidos.

Calificación compuesta. La suma directa o ponderada de las calificaciones en dos o más pruebas o secciones de una prueba.

Calificación cruda. Calificación del examinado en una prueba que no se ha convertido, se calcula como el número de reactivos que se responden correctamente o el número de respuestas correctas menos cierta porción de las respuestas incorrectas.

Calificación de eliminación. Procedimiento de calificación en el cual en lugar de marcar sólo la mejor respuesta para un reactivo, el examinado indica qué opciones son incorrectas.

Calificación derivada. Una calificación que se obtiene al realizar algunas operaciones matemáticas sobre una calificación cruda, como multiplicar la calificación cruda por una constante y/o sumar una constante a la calificación. (Vea *calificaciones estándar, calificaciones T, calificación z.*)

Calificación empírica. Sistema de calificación en el cual las respuestas del examinado se califican de acuerdo con una clave elaborada a partir de las respuestas que dan las personas en ciertos grupos criterio, como los esquizofrénicos o los médicos. Este procedimiento de calificación se emplea con diversos inventarios de intereses y de personalidad.

Calificación equivalente a la edad. Vea *norma de edad.*

Calificación holística. Procedimiento de calificación, como en los reactivos de ensayo, en la cual se asigna una sola calificación en términos del desempeño global del individuo, en lugar de asignar puntos diferentes a características distintas de la respuesta. (Vea *calificación analítica.*)

Calificación límite. Todos los solicitantes que caen por debajo de la calificación límite en un criterio son rechazados, y todos los solicitantes que obtienen una calificación en el límite o encima de éste se aceptan. La calificación límite depende de la validez de la prueba, la razón de selección y otros factores.

Calificación real. La calificación hipotética que es una medida del verdadero conocimiento que tiene el examinado del material de la prueba. En la teoría de los tests, la calificación real de un examinado en una prueba es la media de la distribución de calificaciones que resultaría si el examinado presentara la prueba un número infinito de veces.

Calificación *z*. Cualquiera de un grupo de calificaciones derivadas que varían de $-\infty$ a $+\infty$ calculada de la fórmula $z =$ (calificación cruda – media)/ desviación estándar, para cada calificación cruda. En una distribución normal, más de 99% de los casos cae entre $z = -3.00$ y $z = +3.00$.

Calificaciones estándar normalizadas. Calificaciones obtenidas al transformar las calificaciones crudas de tal manera que las calificaciones transformadas se distribuyen normalmente con una media de 0 y una desviación estándar de 1 (o alguna función lineal de esos números).

Calificaciones estándar. Grupo de calificaciones, como las calificaciones *z*, las calificaciones *T* o las calificaciones estaninas que tienen una media y una desviación

estándar deseadas. Las calificaciones estándar se calculan cambiando las calificaciones crudas a calificaciones *z*, multiplicando las calificaciones *z* por la desviación estándar deseada y luego sumando la media deseada de las calificaciones transformadas al producto.

Calificaciones sumadas (método de). Técnica de elaboración de escalas de actitud elaborada por R. Likert. Los calificadores verifican los valores numéricos en un continuo con tres a siete (por lo general cinco) categorías que corresponden al grado positivo o negativo de cada una de un gran número de afirmaciones de actitud relacionadas con el tema en cuestión. Se seleccionan aproximadamente veinte afirmaciones de acuerdo con ciertos criterios estadísticos para componer la escala final de actitudes.

Calificaciones *T*. Calificaciones estándar convertidas y normalizadas que tienen una media de 50 y una desviación estándar de 10. Las calificaciones *Z* también son calificaciones estándar con una media de 50 y una desviación estándar de 10, pero en contraste con las calificaciones *T* no son normalizadas.

Cargas factoriales. En el análisis factorial, las correlaciones resultantes (pesos) entre pruebas (y otras variables) y los factores extraídos.

CATI. Entrevista telefónica asistida por computadora. El entrevistador lee en voz alta los reactivos de un cuestionario del monitor de una computadora, y las respuestas del sujeto son registradas y analizadas por la computadora. Dependiendo de la respuesta del sujeto, la computadora puede saltarse ciertos reactivos.

Centro de evaluación. Una técnica, usada principalmente en la selección de personal ejecutivo, para evaluar el comportamiento y las características de personalidad de un pequeño grupo de individuos al hacerlos realizar una variedad de tareas en un periodo de unos cuantos días.

CI de Desviación. Coeficiente intelectual (CI) que se obtiene al convertir las calificaciones crudas obtenidas en una prueba de inteligencia a una distribución de calificaciones que tiene una media de 100 y una desviación estándar fija, como 16 para la Stanford-Binet o 15 para los tests de Wechsler.

Clasificación o ranking. El uso de las calificaciones de una prueba para asignar a una persona a una categoría en lugar de otra. Ordenar a un grupo de individuos de acuerdo con su posición juzgada en cierta característica; colocar en orden una lista de características de un individuo de acuerdo con su importancia.

Cociente de desarrollo (CD). Índice, que equivale aproximadamente a la edad mental, que resume el comportamiento de un infante según se evalúa por los programas de desarrollo de Gesell.

Coeficiente intelectual (CI). Una calificación derivada, usada originalmente en la calificación de la Escala de Inteligencia de Stanford-Binet. Una razón CI se calcula dividiendo la *edad mental* (*EM*) del examinado entre su *edad cronológica* (*EC*) y multiplicando el cociente resultante por 100. Un CI de desviación se calcula multiplicando la calificación *z* que corresponde a una calificación cruda en una prueba de inteligencia por la desviación estándar del CI de desviación y agregando 100 al producto.

Coeficiente alfa. Coeficiente de confiabilidad por consistencia interna, apropiado para pruebas compuestas por reactivos dicotómicos o de puntos múltiples; la correlación esperada de una prueba con forma paralela que contiene la misma cantidad de reactivos.

Coeficiente biserial puntual. Coeficiente de correlación calculado entre una variable dicotómica y una variable continua; se deriva del coeficiente de correlación producto-momento.

Coeficiente de confiabilidad. Un índice numérico, entre .00 y 1.00, de la confiabilidad de un instrumento de evaluación. Los métodos para determinar la confiabilidad incluyen test-retest, formas paralelas y consistencia interna.

Coeficiente de consistencia interna. Coeficiente de confiabilidad basado en estimaciones de la consistencia interna de una prueba (por ejemplo, coeficiente de división por mitades y coeficiente alfa).

Coeficiente de correlación. Índice numérico del grado de relación entre dos variables. Los coeficientes de correlación por lo general varían de –1.00 (relación negativa perfecta) a .00 (ausencia total de relación) hasta +1.00 (relación positiva perfecta). Dos tipos comunes de coeficientes de correlación son el coeficiente producto-momento y el coeficiente biserial puntual.

Coeficiente de correlación múltiple (R). Medida del grado total de relación, varía entre –1.00 y +1.00, entre varias variables con una sola variable criterio. La correlación múltiple de un grupo de pruebas de aptitudes académicas con las calificaciones escolares suele ser de alrededor de .60 a .70, un grado moderado de correlación.

Coeficiente de división por mitad. Un estimado de confiabilidad determinado al aplicar la fórmula de Spearman-Brown para $m = 2$ a la correlación entre las dos mitades de la misma prueba, como los reactivos con números nones y los reactivos con números pares.

Coeficiente de equivalencia. Coeficiente de confiabilidad (correlación) obtenido al aplicar dos formas diferentes de una prueba a la misma gente. (Vea *confiabilidad de formas paralelas.*)

Coeficiente de estabilidad. Coeficiente de confiabilidad (correlación) obtenido al aplicar una prueba al mismo

grupo de sujetos en dos ocasiones diferentes. (Vea *confiabilidad de test-retest.*)

Coeficiente de estabilidad y equivalencia. Coeficiente de confiabilidad que se obtiene al aplicar dos formas de una prueba a un grupo de sujetos en dos ocasiones diferentes.

Coeficiente de generalización. Coeficiente numérico que es un indicador del grado de generalización (es decir, confiabilidad) de la muestra a la población. Un coeficiente de generalización toma en cuenta una o más fuentes de error al generalizar de la muestra a la población. Se calcula como una razón de la suma de las varianzas de los componentes de la calificación de la prueba bajo consideración a esta suma, más la suma ponderada de las varianzas de error en la situación.

Coeficiente de validez. La correlación entre las calificaciones en una variable de predicción y las calificaciones en una variable criterio.

Cognición. Tiene que ver con los procesos del intelecto; recordar, pensar, resolver problemas y cosas similares.

Comportamiento adaptado. Grado en que una persona es capaz de interactuar de manera efectiva y apropiada con el ambiente.

Comportamiento no verbal. Cualquier conducta comunicativa que no implica hacer sonidos de palabras o señales. Incluye movimientos de partes corporales grandes (*macrocinestésicas*) y pequeñas (*microcinestésicas*), distancia interpersonal o territorialidad (*proxémica*), tono y tasa de sonidos vocales (*paralingüística*) y comunicaciones impartidas por los asuntos prescritos por la cultura relacionados con el tiempo, el vestuario, la pertenencia, etcétera.

Comportamientos objetivo. Conductas específicas, definidas de manera objetiva observadas y medidas en las evaluaciones conductuales. De particular interés son los efectos que los eventos antecedentes y consecuentes tienen en esas conductas.

Comunalidad. Proporción de varianza en una variable medida explicada por la varianza que la variable tiene en común con otras variables.

Comunicación privilegiada. Comunicación confidencial entre una persona y su abogado, doctor, pastor o cónyuge. La información comunicada es privilegiada y no puede revelarse en la corte si la persona (cliente, paciente, penitente, cónyuge) reclama el privilegio.

Confiabilidad. El grado en que un instrumento de evaluación psicológica mide algo en forma consistente. Un instrumento confiable está relativamente libre de errores de medición, por lo que las calificaciones obtenidas en el instrumento son cercanas en valor numérico a las verdaderas calificaciones de los examinados.

Confiabilidad de formas paralelas. Índice de confiabilidad (coeficiente de equivalencia) que se determina correlacionando las calificaciones de los individuos en forma paralela de una prueba con sus calificaciones.

Confiabilidad de test-retest. Método para evaluar la confiabilidad de una prueba aplicándola al mismo grupo de examinados en dos ocasiones diferentes y calculando la correlación (coeficiente de estabilidad) entre sus calificaciones en las dos ocasiones.

Confiabilidad entre calificadores. Dos calificadores asignan una calificación numérica a una muestra de personas. Luego se calcula la correlación entre los dos conjuntos de números.

Confiabilidad impar-par. La correlación entre las calificaciones totales en los reactivos con número impar y las calificaciones totales en los reactivos con número par de una prueba, corregida mediante la fórmula de confiabilidad de Spearman-Brown. (Vea *fórmula de la profecía Spearman-Brown.*)

Confiabilidad intraclase. Índice de acuerdo entre las calificaciones asignadas por un grupo de calificadores ("jueces") a una característica o conducta de una persona.

Confiabilidad por concordancia. Varios calificadores hacen juicios numéricos de la cantidad de una característica o conducta mostrada por una muestra grande de gente. Luego se calcula un coeficiente de concordancia, un índice de acuerdo entre los juicios de los calificadores.

Consentimiento informado. Acuerdo formal que establece una persona, o su tutor o representante legal, con un organismo o con alguien más para permitir el uso del nombre de la persona y/o información personal (calificaciones de la prueba y similares) para un propósito especificado.

Consistencia interna. El grado en el cual todos los reactivos de una prueba miden la misma variable o constructo. La confiabilidad de una prueba calculada por las fórmulas Spearman-Brown, Kuder-Richardson o alfa de Cronbach es una medida de la consistencia interna de la prueba.

Constructo. Variable o concepto que una prueba está diseñada para medir.

Contaminación de criterios. El efecto de cualquier factor sobre un criterio de modo que el criterio no es una medida válida de los logros de un individuo. Las puntuaciones en una prueba de aptitud pueden utilizarse para predecir las calificaciones en la escuela, pero cuando los profesores usan las puntuaciones de una prueba de aptitud para decidir qué calificaciones asignar a los estudiantes, las calificaciones no son un criterio válido para validar la prueba de aptitud; el criterio ha sido contaminado.

Corrección para la adivinación. Fórmula que se aplica a las puntuaciones crudas para corregir los efectos de la adivinación aleatoria por parte de los examinados. Una fórmula popular de corrección para la adivinación requiere que se reste parte del número de reactivos que el examinado responde incorrectamente del número que responde correctamente.

Corrección para la atenuación. Fórmula utilizada para estimar cuál sería el coeficiente de validez de una prueba si tanto la prueba como el criterio fueran totalmente confiables.

Correlación. Grado de relación o asociación entre dos variables, tales como una prueba y una medida de criterio.

Crecimiento esperado. Cambio promedio en las calificaciones de una prueba que ocurren a lo largo de un periodo determinado en personas de una edad, grado u otras características especificadas.

Criterio. Estándar o variable con la cual se comparan las calificaciones obtenidas en un instrumento psicométrico o contra la cual se evalúa. La validez de una prueba u otro procedimiento psicométrico usado para seleccionar o clasificar a la gente es determinada por su capacidad para predecir un criterio especificado de conducta en la situación para la cual se seleccionan o clasifican las personas.

Cuartil. Calificación en una distribución de frecuencia por debajo del cual cae ya sea el 25% (primer cuartil), 50% (segundo cuartil), 75% (tercer cuartil) o 100% (cuarto cuartil) del número total de calificaciones.

Curva (característica) de respuesta a los reactivos. Gráfica que muestra la proporción de individuos que responden correctamente a un reactivo de la prueba, graficado contra un criterio interno (la calificación total en la prueba) o externo de desempeño.

Curva característica de los reactivos. Gráfica utilizada en el análisis de reactivos, en la cual la proporción de examinados que pasa un reactivo específico se grafica contra las calificaciones totales de la prueba.

Decil. Uno de los nueve puntos de calificación que dividen una distribución de calificación en diez partes iguales.

Demencia. Término legal para un trastorno del juicio o el comportamiento en el cual una persona no puede distinguir entre el bien y el mal (*regla McNaghten*) o no puede controlar o manejar sus acciones o asuntos.

Demencia vascular. El tipo más común de demencia en la vejez; se debe a una enfermedad cerebrovascular asociada con hipertensión y daño vascular del cerebro.

Dependencia de campo. Un estilo perceptual en el cual el receptor depende principalmente de señales del ambiente visual circundante, más que de señales cinestésicas (gravitacionales), para determinar la posición hacia arriba en la prueba de varilla y marco.

Desviación estándar. La raíz cuadrada de la varianza; se usa como una medida de la dispersión o extensión de un grupo de calificaciones.

Detección. Término general para cualquier proceso de selección, por lo general no muy preciso, mediante el cual algunos solicitantes son aceptados y otros son rechazados.

Determinantes específicos. Palabras como *todos, siempre, nunca* y *sólo*, que indican que un reactivo de verdadero-falso probablemente es falso, o *a menudo, en ocasiones* y *usualmente*, que sugieren que la afirmación del reactivo es verdadera.

Diagrama de dispersión. Grupo de puntos graficados a partir de un conjunto de valores de X, Y, donde X es la variable independiente y Y la variable dependiente.

Diferencial semántico. Escala de calificación para evaluar los significados connotativos de conceptos seleccionados. Cada concepto se califica en una escala bipolar de adjetivos de siete puntos.

Discalculia. Incapacidad para realizar operaciones aritméticas.

Discapacidad de aprendizaje. Dificultad para aprender a leer, escribir, deletrear o realizar operaciones aritméticas y otras habilidades académicas observada en una persona cuya calificación en una prueba de inteligencia (CI) es promedio o más alta.

Discapacidad específica de aprendizaje. (Vea *discapacidad de aprendizaje.*)

Discusión de grupo sin líder (LDG). Seis o más individuos (por ejemplo, candidatos a un puesto ejecutivo) son observados mientras discuten un problema asignado para determinar su efectividad para trabajar en equipo y llegar a una solución.

Dislexia. Trastorno de la lectura asociado con un deterioro en la habilidad para interpretar relaciones espaciales o para integrar información auditiva y visual.

Distractores. Cualquiera de las opciones incorrectas en un reactivo de opción múltiple.

Distribución bimodal. Una distribución de frecuencia que tiene dos modas (puntos máximos). (Vea *distribución de frecuencia* y *moda.*)

Distribución de frecuencia. Tabla de intervalos de calificaciones y el número de casos (calificaciones) que caen dentro de cada intervalo.

Distribución normal. Distribución de frecuencias lisa y acampanada, simétrica con respecto de la media y descrita por una función matemática exacta. Las calificaciones de prueba de un grupo grande de sujetos con frecuencia se distribuyen aproximadamente de esta manera.

Ectomorfo. En el sistema de somatotipos de Sheldon, una persona con constitución corporal alta y delgada;

relacionado con el tipo de temperamento cerebrotónico (pensamiento, introvertido).

Ecuación de regresión. Ecuación lineal para predecir calificaciones criterio a partir de las calificaciones en una o más variables predictoras; un procedimiento que a menudo se utiliza en los programas de selección o diagnóstico y predicción actuarial.

Edad basal. El nivel de año más alto en una prueba de inteligencia, como en las antiguas ediciones de la Stanford-Binet, por debajo del cual el examinado pasa todas las subpruebas.

Edad del desarrollo. Calificación en los programas de desarrollo de Gesell.

Edad mental (EM). Calificación derivada en una prueba de inteligencia como la de Stanford-Binet. La edad mental del individuo corresponde a la edad cronológica de una muestra representativa de niños de la misma edad cronológica cuya calificación promedio en la prueba es igual a la calificación del examinado. (Vea *coeficiente intelectual.*)

Edad tope. La edad o nivel de año mínimos en una prueba, como la Stanford-Binet, en la cual el examinado falla todas las subpruebas. (Vea *edad basal.*)

Efecto Barnum. Aceptar como exacta una descripción de personalidad planteada en generalidades, perogrulladas y otras afirmaciones que parecen específicas de una persona determinada pero que en realidad pueden aplicarse casi a cualquier individuo. Es igual al *error de la tía Fanny*.

Efecto de halo. Asignar a una persona una calificación alta en una característica sólo porque obtiene calificaciones altas en otras características.

Efecto de las expectativas. Efecto de las expectativas de un maestro sobre las calificaciones CI de sus pupilos; de manera más general, el efecto de las expectativas de una persona sobre el comportamiento de otro individuo.

Efecto de primacía. La tendencia de los calificadores a asignar más peso a las conductas o desempeños iniciales que a las conductas subsecuentes de los calificados.

Ego (YO). De acuerdo con la teoría psicoanalítica, la parte de la personalidad (el "yo") que obedece el principio de realidad e intenta mediar en el conflicto entre el *id* y el superego.

Electroencefalógrafo (EEG). Aparato electrónico diseñado para detectar y registrar las ondas cerebrales del cuero cabelludo intacto.

Electromiógrafo (EMG). Aparato electrónico diseñado para medir la actividad o tensión muscular.

Encadenamiento. Metodología basada en las respuestas a los reactivos para igualar dos pruebas transformando los parámetros de los reactivos de una forma de la prueba a los de la segunda forma del instrumento, de forma que los parámetros correspondientes en las dos pruebas estarán en la misma escala numérica.

Encuesta de opinión. Preguntar a una muestra de una población objetivo sobre sus opiniones concernientes a objetos, temas y eventos particulares.

Endomorfo. En el sistema de somatotipos de Sheldon, una persona que tiene una forma corporal robusta (obeso); se relaciona con el temperamento viscerotónico (relajado, sociable).

Enfermedad de Alzheimer. Síndrome cerebral crónico, que usualmente ocurre en la vejez, caracterizado por el deterioro gradual de la memoria, desorientación y otras características de demencia.

Entrevista. Procedimiento sistemático para obtener información al plantear preguntas y, en general, al interactuar verbalmente con una persona (el entrevistado).

Entrevista de diagnóstico. Entrevista diseñada para obtener información sobre los pensamientos, sentimientos, percepciones y comportamiento de una persona; se utiliza para tomar una decisión de diagnóstico acerca de la persona.

Entrevista de estrés. Procedimiento de entrevista en el cual el entrevistador aplica técnicas psicológicamente estresantes (cuestionamiento crítico y hostil, interrupciones frecuentes, silencios prolongados, etc.) para romper las defensas del entrevistado y/o para determinar cómo reacciona el entrevistado bajo presión.

Entrevista estructurada. Procedimiento de entrevista en el cual al entrevistado se le plantea un conjunto predeterminado de preguntas.

Equilibrio. En la teoría de J. Piaget del desarrollo cognoscitivo, el proceso por el cual un niño llega a conocer y a entender el ambiente al interactuar con él. El equilibrio involucra los procesos de asimilación y acomodación.

Equivalentes a la curva normal (NCEs). Calificaciones estándar normalizadas que tienen una media de 50 y una desviación estándar de 21.06 que va de 1 a 99.

Error de contraste. Al entrevistar o calificar, la tendencia a evaluar a una persona de manera más positiva si el individuo anterior recibió una evaluación sumamente negativa, o evaluar a una persona de forma más negativa si la persona anterior recibió una evaluación muy positiva.

Error de indulgencia. Tendencia a calificar a un individuo a un nivel más alto en una característica positiva y de manera menos severa en una característica negativa de lo que debería ser calificado.

Error de la tía Fanny. Aceptar como exacta una descripción de la personalidad trivial y muy generalizada que podría pertenecer casi a cualquier persona, incluso a la Tía Fanny.

Error de medición. La diferencia entre una calificación observada y la calificación verdadera correspondiente en una prueba.

Error estándar de estimación. Grado de error cometido al estimar la calificación de una persona en una variable criterio a partir de su calificación en una variable predictora.

Error estándar de medición. Estimación de la desviación estándar de la distribución normal de las calificaciones de la prueba que un examinado debería obtener en teoría al presentar la prueba un número infinito de veces. Si la calificación obtenida por el examinado en la prueba es *X*, entonces las probabilidades son dos de tres de que forme parte de un grupo de personas cuyas verdaderas calificaciones en la prueba caen dentro de un error estándar de medición de X.

Error fundamental de atribución. Tendencia a atribuir la conducta propia a las influencias de la situación, pero atribuir el comportamiento de otra gente a factores disposicionales.

Escala. Sistema de números graduados, usados al asignar valores medidos a características seleccionadas de objetos, eventos o personas.

Escala de actitudes. Instrumento de lápiz y papel que consiste en una serie de afirmaciones concernientes a una institución, situación, persona o evento. El examinado responde a cada afirmación al confirmarla o indicar su grado de acuerdo o desacuerdo con ella.

Escala de calificación. Lista de palabras o afirmaciones concernientes a rasgos o características, en ocasiones en la forma de una línea continua dividida en secciones correspondientes a los grados de las características, en las cuales el calificador indica los juicios de su propio comportamiento o características o del comportamiento o características de otra persona (calificado). El calificador indica cómo o en qué grado la conducta o característica es poseída por el sujeto.

Escala de calificación gráfica. Escala de calificación que contiene una serie de reactivos, cada uno de los cuales consiste en una línea sobre la cual el calificador hace una marca para indicar el grado de una característica que percibe que posee el individuo. Por lo general, en el extremo izquierdo de la línea hay una breve descripción verbal que indica el menor grado de la característica, y en el extremo derecho hay una descripción del grado más alto de la característica. También pueden encontrarse descripciones breves de los grados intermedios de la característica en puntos equidistantes a lo largo de la línea.

Escala de edad. Una prueba en la cual los reactivos se agrupan por nivel de edad.

Escala de hombre a hombre. Procedimiento en el cual las calificaciones en un rasgo específico (por ejemplo, liderazgo) se asignan al comparar a cada persona que va a calificarse con otras personas cuya posición en el rasgo ya se determinó.

Escala de intervalo Una escala de medición en la cual la igualdad de las diferencias numéricas implica igualdad de las diferencias en el atributo o característica que se mide. Las escalas de temperatura (Celsius, Fahrenheit) y supuestamente las escalas de calificación estándar (*z*, *T*, etc.) son ejemplos de escalas de intervalo.

Escala de Likert. Escala de actitudes en la cual los individuos indican su grado de acuerdo o desacuerdo con una proposición particular concerniente a algún objeto, persona o situación.

Escala de puntos. Prueba en la cual se asignan puntos (por ejemplo, 0, 1 o 2) a cada reactivo, dependiendo de lo precisa y completa que sea la respuesta.

Escala de razón. Escala de medición, que tiene un verdadero cero, en la cual razones numéricas iguales implican razones iguales del atributo que se mide. Las variables psicológicas por lo general no se miden en escalas de razón, lo que sí sucede con la estatura, peso, energía y muchas otras variables físicas.

Escala nominal. Nivel inferior de medición, en el cual los números se utilizan sólo como descriptores o nombres de las cosas, en lugar de designar orden o cantidad.

Escala ordinal. Tipo de escala de medición en la cual los números se refieren meramente a los rangos de objetos o eventos arreglados en orden de mérito (por ejemplo, los números que se refieren al orden de terminación en un concurso).

Escala TC (tomografía computarizada). Procedimiento de diagnóstico basado en los rayos X en el cual se genera mediante la computadora una representación tridimensional del cerebro.

Escala visual análoga. Instrumento psicométrico para la medición de experiencias subjetivas como el dolor, la ansiedad y los anhelos por ciertas sustancias. El paciente señala o marca el punto en una línea que corresponde a la intensidad de su experiencia.

Especificidad. La proporción de la varianza total de una prueba que se debe a factores específicos de la prueba misma.

Esquema. En la teoría de J. Piaget del desarrollo cognoscitivo, una estructura mental (asir, succionar, sacudir, etc.) que es modificada (acomodada) como resultado de la experiencia.

Estadístico. Número que se utiliza para describir algunas características de una muestra de calificaciones de la prueba, como la media aritmética o la desviación estándar.

Estandarización. Aplicar una prueba cuidadosamente elaborada a una muestra grande y representativa de personas bajo condiciones estándar con el propósito de determinar las normas.

Estaninas. Escala de calificación estándar que consta de calificaciones de 1 al 9, las calificaciones estaninas tienen una media de 5 y una desviación estándar aproximadamente de 2.

Estilo cognoscitivo. Estrategia o aproximación a la percepción, memoria y pensamiento que una persona parece preferir al tratar de entender y enfrentar el mundo (por ejemplo, independencia-dependencia del campo, reflexión-impulsividad y locus de control interno-externo).

(Estilo de) grupo de respuestas de consentimiento. Tendencia que presenta una persona a responder afirmativamente ("sí" o "verdadero") a los reactivos de las pruebas de personalidad y en otras situaciones alternativas de respuesta.

Estrategia de instigación graduada. Procedimiento dinámico de evaluación en el cual el examinador presenta una serie de indicios conductuales para enseñar las reglas que se necesitan para completar con éxito una tarea de prueba. Los indicios o instigaciones, los cuales se generan de un guión predeterminado más que de las respuestas del sujeto, son bastante generales al inicio, pero se vuelven más específicos conforme se necesita.

Estudio de caso. Estudio detallado de un individuo, diseñado para proporcionar una comprensión a fondo de la conducta y la personalidad. La información para un estudio de caso se obtiene de datos biográficos, de entrevista, de observaciones y de pruebas.

Etapa de fantasía. La etapa más temprana en el desarrollo de los intereses, en la cual las orientaciones de los intereses de un niño no se basan en una percepción exacta de la realidad.

Etapa de las operaciones concretas. En la teoría de J. Piaget del desarrollo cognoscitivo, la etapa (de los 7 a los 11 años) durante la cual un niño desarrolla sistemas organizados de operaciones por medio del proceso de interacción social, con una reducción correspondiente en el egocentrismo.

Etapa preoperacional. En la teoría de J. Piaget del desarrollo cognoscitivo, el periodo egocéntrico del desarrollo (de los tres a los siete años) cuando el niño adquiere el lenguaje y otras representaciones simbólicas.

Etapa realista. Etapa final en el desarrollo de los intereses vocacionales, que por lo general ocurre al final de la adolescencia o al inicio de la adultez. En esta etapa el individuo tiene una idea realista acerca de qué ocupaciones particulares implica la vocación que le gustaría seguir.

Etapa sensoriomotriz. La primera etapa en la teoría de J. Piaget del desarrollo cognoscitivo (de 0 a 2 años), durante la cual el niño aprende a ejercitar reflejos simples y a coordinar varias percepciones.

Etiqueta menos estigmatizante. Etiqueta o categoría de clasificación que se considera que es la que implica menor estigma social y por ende es apropiada para la condición diagnosticada.

Evaluación. Juzgar el mérito o valor del comportamiento de un individuo a partir de una combinación de calificaciones de prueba, observaciones e informes.

Evaluación afectiva. Medición de variables o características no cognoscitivas (no del intelecto). Las variables afectivas incluyen temperamento, emociones, intereses, actitudes, estilo personal y otros comportamientos, rasgos o procesos típicos de un individuo. (Vea *evaluación cognoscitiva.*)

Evaluación auténtica. Evaluación del desempeño en tareas realistas o de la vida real o en situaciones reales.

Evaluación automatizada. Uso de máquinas para calificar las pruebas, computadoras y otros aparatos electrónicos o electromecánicos para aplicar, calificar e interpretar las evaluaciones psicológicas.

Evaluación cognoscitiva. Medición de los procesos intelectuales, como la percepción, memoria, pensamiento, juicio y razonamiento. (Vea *evaluación afectiva.*)

Evaluación de desempeño. Tipo de procedimiento de evaluación que requiere que los estudiantes construyan, creen o demuestren algo. En la mayoría de los casos, hay muchas maneras de evaluar el desempeño y más de una respuesta aceptable.

Evaluación de personalidad. Descripción y análisis de la personalidad por medio de diversas técnicas, incluyendo la observación y la entrevista, la aplicación de listas de verificación, escalas de calificación, inventarios de personalidad y técnicas proyectivas.

Evaluación dinámica. Aproximación a la evaluación de prueba-enseñanza-prueba en la cual se evalúa a una persona (se le aplica un pretest), luego se le proporciona práctica en los materiales de la prueba y por último se le aplica la prueba de nuevo (el postest). El cambio en el nivel de desempeño del pretest al postest es una medida del potencial de aprendizaje. (Vea *zona de desarrollo potencial.*)

Evaluación formativa. Evaluación del desempeño con el propósito de mejorar la instrucción o determinar áreas de fortaleza y debilidad para el enriquecimiento o la instrucción de remedio (Vea *evaluación sumatoria.*)

Evaluación neuropsicológica. Medición del desempeño cognoscitivo, perceptual y motriz para determinar la localización, grado y efectos del daño y disfunción neurológicas.

Evaluación sumatoria. Evaluación al final de una unidad de instrucción o curso de estudio para proporcionar una suma total o medida del producto final del aprovechamiento.

Extrovertido. Término de C. G. Jung para referirse a los individuos que poseen una orientación social o de pensa-

miento hacia el entorno externo y las otras personas, más que hacia sus propios pensamientos y sentimientos.

Factor. Dimensión, rasgo o característica de personalidad revelada al realizar un análisis factorial de la matriz de correlaciones calculada a partir de las calificaciones de un gran número de personas en varias pruebas o reactivos diferentes.

Factor g. El factor general único de inteligencia propuesto por Charles Spearman para explicar las correlaciones elevadas entre las pruebas de inteligencia.

Falometría. La medición de la respuesta eréctil del varón como una medida científica de las preferencias sexuales de los hombres.

Falso positivo. Error de selección o error de decisión de diagnóstico en el cual un procedimiento de evaluación predice de manera incorrecta un resultado adaptado (por ejemplo, alto aprovechamiento, buen desempeño o ausencia de psicopatología).

Falso negativo. Error de selección o error de decisión de diagnóstico en el cual un procedimiento de evaluación predice de manera incorrecta un resultado inadaptado (por ejemplo, bajo rendimiento, mal desempeño o psicopatología).

Fidelidad. La parte angosta de la anchura de banda de una prueba u otro instrumento de medición. Una prueba con alta fidelidad hace una buena predicción de un rango de criterios bastante angosto. (Vea *anchura de banda.*)

Formas equivalentes. (Ver *formas paralelas.*)

Formas paralelas. Dos pruebas que son equivalentes en el sentido de que contienen los mismos tipos de reactivos de igual dificultad y están altamente correlacionadas. Las calificaciones obtenidas en una forma de la prueba son muy cercanas a las obtenidas por las mismas personas en la otra forma.

Fórmula de calificación. Fórmula utilizada para calcular las calificaciones crudas de una prueba. Algunas fórmulas de calificación comunes son $S = R$ y $S = R - W/(k\text{-}1)$; S es la calificación, R es el número de aciertos; W es el número de errores y k es el número de opciones por reactivo.

Fórmula de profecía Spearman-Brown. Fórmula general para estimar la confiabilidad (r_{11}) de una prueba en la cual el número de reactivos es incrementado por un factor de m. En la fórmula $r_{mm} = mr_{11}/[1 + (m - 1)\, r_{11}]$, r_{11} es la confiabilidad de la prueba original (no aumentada) y m es el factor por el cual es aumentada.

Fórmulas de Kuder-Richardson. Fórmulas usadas para calcular una medida de confiabilidad por consistencia interna a partir de una sola aplicación de una prueba con una calificación de 0 a 1.

Frenología. Teoría y práctica desacreditadas que relacionan las características afectivas y cognoscitivas con la configuración (protuberancias) del cráneo.

Funcionamiento diferencial de un reactivo (FDR). Un reactivo de una prueba es más fácil o más discriminante en un grupo que en otro.

Generalización de la validez. La aplicación de la evidencia de la validez a situaciones distintas a aquellas en las que se obtuvo la evidencia.

Grafología. Análisis de la escritura para determinar el carácter o personalidad de la persona que escribe.

Grupo de norma. Muestra de personas en las cuales se estandariza una prueba.

Grupo de respuestas deseables para la sociedad. Tendencia o estilo de respuesta que afecta las calificaciones en un instrumento de evaluación psicológica. Se refiere a la tendencia a que un examinado responda en la dirección que percibe que es la más deseable socialmente, en lugar de responder de manera que sea verdaderamente característica o descriptiva de él.

Grupos de respuesta (estilos). Tendencias de los individuos a responder de maneras relativamente fijas o estereotipadas en situaciones en las que existen dos o más opciones de respuesta, como en los inventarios de personalidad. Las tendencias a adivinar, a responder que es verdadero (conformidad) y a dar respuestas socialmente deseables son algunos de los grupos de respuesta que han sido investigados.

Habilidad verbal. Entender las palabras y sus interrelaciones en el lenguaje; la capacidad para resolver problemas relacionados con palabras.

Habilidades psicomotrices. Habilidades que involucran actividades motoras, como lanzar, atrapar, insertar y manipular objetos de alguna otra manera (por ejemplo, habilidades atléticas).

Hipótesis de transferencia (diferenciación de capacidades). Hipótesis de G. A. Ferguson de que las diferentes capacidades aisladas por el análisis factorial son el resultado del aprendizaje excesivo y de la transferencia diferencial positiva en ciertas áreas de aprendizaje.

IDEA. Acta de los individuos con discapacidades.

Identificación. Asumir las características personales de otra persona, como cuando un niño en desarrollo se identifica con otra persona significativa. También, en la teoría psicoanalítica, un mecanismo de defensa del ego para enfrentar la ansiedad.

IEP. Plan individualizado de educación; consiste en objetivos educativos a corto y a largo plazo para un estu-

diante particular, por lo general con discapacidad de aprendizaje, y los procedimientos para alcanzarlos.

Imagenología por resonancia magnética (escáner MRI). Procedimiento de diagnóstico en el cual una computadora traza los cambios en la resonancia magnética de los átomos en el cerebro.

Impacto adverso. Situación en la cual la tasa de selección es sustancialmente menor para los miembros de una raza, sexo o grupo étnico que para los de otro.

Incidente crucial. Comportamiento que se considera crucial para el desempeño efectivo en un trabajo (por ejemplo, "limpia el área de trabajo antes de partir" o "trata a los clientes con cortesía").

Independencia de campo. Un estilo perceptual en el cual el perceptor depende sobre todo de señales cinestésicas (gravitacionales), más que de las señales visuales del ambiente circundante, para determinar la posición hacia arriba en la prueba de varilla y marco.

Índice de dificultad de reactivos. Índice de la facilidad o dificultad de un reactivo para un grupo de individuos. Una medida conveniente de la dificultad de un reactivo es el porcentaje (p) de examinados que selecciona la respuesta correcta.

Índice de discriminación de reactivos. Medida de la efectividad con que un reactivo discrimina entre los examinados que obtienen una calificación alta en la prueba como un todo (o en alguna otra variable criterio) y los que obtienen bajas calificaciones.

Índice de herencia (h^2). Razón entre la varianza de la calificación de una prueba atribuible a la herencia y la varianza atribuible a la combinación de la herencia y el entorno.

Inteligencia. Se han ofrecido muchas definiciones de este término, como "la capacidad para juzgar, entender y razonar adecuadamente" (A. Binet) y "la capacidad para el pensamiento abstracto" (L. M. Terman). En general, lo que miden las pruebas de inteligencia es la habilidad para tener éxito en tareas de tipo escolar.

Inteligencia cristalizada. Término de R. B. Cattell para la capacidad mental (conocimiento, habilidades) adquirida mediante la experiencia y la educación.

Inteligencia experiencial. De acuerdo con Sternberg, la habilidad para enfrentar de manera efectiva las tareas nuevas.

Inteligencia fluida. Término de R. B. Cattell para referirse a la capacidad mental inherente que se determina de manera genética, como se ve en la solución de problemas o las respuestas novedosas.

Intervalo de confianza. Rango de valores dentro del cual podemos estar casi seguros (por lo general una confianza de 95% a 99%) de que cae la verdadera calificación de una persona (o la diferencia entre calificaciones) en una prueba o un criterio variable. (Vea *error estándar de medición* y *error estándar de estimación.*)

Intervalos de igual aparición (método de). Método de escalamiento de actitudes desarrollado por L. L. Thurstone en el cual una muestra grande de "jueces" clasifica afirmaciones de actitudes en 11 pilas de acuerdo con lo favorable de la actitud expresada en la afirmación. El valor de la escala de una afirmación se calcula como la mediana y el índice de ambigüedad como el rango semi-intercuartil de las calificaciones de los jueces.

Introvertido. Término de Carl Jung para la orientación hacia el yo; interés principal por los pensamientos y sentimientos propios, más que por el ambiente externo o por otras personas; preferencia por actividades solitarias.

Inventario. Conjunto de preguntas o afirmaciones a las que responde el individuo (por ejemplo, indicando su acuerdo o desacuerdo), diseñado para proporcionar una medida de la personalidad, intereses, actitudes o comportamiento.

Inventario de autorreporte. Prueba de lápiz y papel de los rasgos de personalidad o intereses, compuesta por una serie de reactivos que el examinado señala como característicos (verdaderos) o no característicos (no verdaderos) de sí mismo.

Inventario de intereses. Una prueba o lista de verificación, como el Inventario de Intereses de Strong o el Estudio de Intereses Generales de Kuder, diseñado para evaluar las preferencias del individuo por ciertas actividades y temas.

Inventario de personalidad. Inventario de autorreporte o cuestionario que consta de afirmaciones concernientes a características y conductas personales. En un inventario de verdadero-falso, el sujeto indica si cada reactivo lo describe; en un inventario de opción múltiple o elección forzada, el individuo selecciona las palabras, frases o afirmaciones que lo describen.

Justicia. En un test de aptitudes, el grado en que las calificaciones no presentan sesgos, es decir, son igualmente predictivas del desempeño criterio de diferentes grupos.

Límites múltiples. Estrategia de selección en la cual se requiere que los solicitantes obtengan al menos las calificaciones mínimas específicas en varios criterios de selección para ser aceptados (contratados, admitidos, etcétera).

Lóbulo frontal. Parte de la corteza cerebral en los lóbulos frontales anteriores a la fisura central.

Lóbulo occipital. Área de la corteza cerebral que se encuentra en la parte trasera de la cabeza; es especialmente importante para la visión.

Lóbulo parietal. Parte de la corteza cerebral; se localiza detrás de la fisura central y entre los lóbulos frontal y

occipital; contiene estructuras nerviosas para experimentar sensaciones somestésicas.

Lóbulos temporales. Lóbulos del cerebro localizados en la región temporal de la cabeza. Juegan un papel importante en la audición, habla y otras funciones neurológicas de orden superior.

Locus de control. Término de J. B. Rotter para un estilo cognoscitivo-perceptual caracterizado por la dirección típica (interna o del yo frente a externa o de otros) desde la cual los individuos perciben que son controlados.

Matriz de rasgos y métodos múltiples. Matriz de coeficientes de correlación que resulta de correlacionar las medidas del mismo rasgo por medio del mismo método, de rasgos distintos por el mismo método, del mismo rasgo por métodos diferentes y de diferentes rasgos por métodos distintos. Las magnitudes relativas de los cuatro tipos de correlaciones se comparan al evaluar la validez de constructo de una prueba.

Media aritmética. Medida del promedio o tendencia central de un grupo de calificaciones. La media aritmética se calcula dividiendo la suma de las calificaciones entre el número de calificaciones.

Mediación. Estrategia de evaluación dinámica en la cual el examinador interactúa continuamente con el examinado para incrementar la probabilidad de que se encuentre una solución a un problema presentado.

Mediana. Punto de calificación en una distribución de calificaciones por debajo y por encima de la cual cae el 50% de las calificaciones.

Medición ipsativa. Formato de reactivos de una prueba (por ejemplo, elección forzada) en el cual las variables medidas se comparan entre sí, de modo que la calificación de una persona en una variable es afectada por sus calificaciones en otras variables medidas por el instrumento.

Medición. Procedimientos para determinar (o indicar) el monto o cantidad de algún constructo o entidad; asignación de números a objetos o eventos.

Memoria explícita. Memoria intencional, consciente.

Memoria implícita. Memoria que ocurre sin intento consciente de recordar.

Mesomorfo. Término de W. H. Sheldon para una persona que tiene un físico atlético; se correlaciona con el temperamento somatotónico (activo, agresivo, enérgico).

Método de observación. Observar la conducta en una situación controlada o no controlada y hacer un registro formal o informal de las observaciones.

Método equipercentil. Método tradicional de convertir las unidades de calificación de una prueba en las unidades de calificación de una prueba paralela. Las calificaciones en cada prueba se convierten en rangos percentilares, y se produce una tabla de calificaciones equivalentes al igualar la calificación en el percentil *p* en la primera prueba al percentil *p* en la segunda prueba.

Moda. La calificación que ocurre con mayor frecuencia en un grupo de calificaciones.

Modelo de Rasch. Modelo de un parámetro (dificultad del reactivo) para ordenar en escala los reactivos de la prueba con propósitos de analizar los reactivos y estandarizar la prueba. El modelo se basa en la suposición de que los índices de adivinación y discriminación de reactivos son parámetros insignificantes. Como sucede con otros modelos de rasgos latentes, el modelo de Rasch relaciona el desempeño del examinado en los reactivos de la prueba (porcentajes de aprobación) con su posición estimada en un rasgo hipotético de habilidad latente o continuo.

Modelo jerárquico. Modelo de árbol de la inteligencia propuesto por P. E. Vernon, consistente en un factor general al nivel más alto, dos factores de grupo principales (verbal-educativo y práctico-mecánico-espacial) en el segundo nivel, y una serie de factores de grupo menores en el tercer nivel.

Modelo pluralista. En el Sistema de Evaluación Pluralista Multicultural (SOMPA), una combinación constituida por los Materiales de Evaluación del Estudiante y la Entrevista con los Padres. La calificación de un niño en las diversas medidas se interpreta comparándola con las calificaciones de otros niños que tienen antecedentes socioculturales similares.

Modelo RIASEC. Modelo de J. L. Holland de tipos persona-ambiente intereses-personalidad que constan de temas realistas, de investigación, artísticos, sociales, emprendedores y convencionales.

Muestra aleatoria. Muestra de observaciones (por ejemplo, calificaciones en una prueba) que se toma de una población de tal manera que cada miembro de la población objetivo tiene la misma oportunidad de ser seleccionado en la muestra.

Muestra de estandarización. Subconjunto de una población objetivo en el cual se estandariza una prueba.

Muestra representativa. Grupo de individuos cuyas características son similares a las de la población de individuos hacia la cual se dirige una prueba.

Muestreo aleatorio estratificado. Procedimiento de muestreo en el cual la población se divide en estratos (por ejemplo, hombres y mujeres; negros y blancos; clase baja, clase media y clase alta), y se seleccionan muestras al azar de los estratos; los tamaños de las muestras dentro de los estratos son proporcionales a los tamaños de los estratos.

Muestreo de reactivos. Procedimiento para seleccionar los subconjuntos de reactivos del grupo total de reactivos; diferentes muestras de reactivos se aplican a diferentes grupos de examinados.

Muestreo incidental. En contraste con el *muestreo de tiempo*, procedimiento de observación en el cual cierto tipo de incidentes, como los que indican un comportamiento agresivo, se seleccionan para su observación y registro.

Muestreo por grupos. Procedimiento de muestreo en el cual la población objetivo se divide en secciones o grupos. El número de unidades seleccionadas al azar de un grupo determinado es proporcional al número total de unidades en el grupo.

Niño excepcional. Un niño que se desvía significativamente del promedio en las características mentales, físicas o emocionales.

Niños especiales. Niños que tienen problemas físicos, psicológicos, cognoscitivos o sociales que hacen que la satisfacción de sus necesidades y potenciales sea más difícil que para otros niños.

Norma de edad. Calificación media obtenida por los niños de una determinada edad cronológica en una prueba de aptitud o aprovechamiento.

Norma de grado (calificación equivalente a la puntuación). El promedio de las calificaciones obtenidas en una prueba por un grupo de niños de un nivel determinado.

Norma según la raza. Basar las normas de calificación de una prueba sólo en una raza o grupo étnico específico, y evaluar las calificaciones de los solicitantes de ese grupo sólo con respecto a esas calificaciones.

Normas. Lista de calificaciones y los correspondientes rangos percentilares, calificaciones estándar u otras calificaciones transformadas de un grupo de personas en las cuales se estandarizó una prueba.

Normas locales. Rangos percentilares, calificaciones estándar u otras normas correspondientes a las calificaciones crudas de una prueba de un grupo local, relativamente pequeño, de examinados.

Normas nacionales. Rangos percentilares, calificaciones estándar u otras normas basadas en una muestra nacional. (Vea *normas locales; normas.*)

Normas percentilares. Lista de calificaciones crudas y los porcentajes correspondientes del grupo de estandarización de la prueba cuyas calificaciones caen por debajo del percentil determinado.

Observación participante. Técnica de investigación, usada sobre todo por los antropólogos culturales, en la cual un observador intenta minimizar lo intrusivo de su persona y actividades de observación convirtiéndose en parte del grupo o situación que está siendo observada, por ejemplo, al vestirse y actuar como las otras personas en el grupo o situación.

Observaciones que no interfieren. Observaciones hechas sin interferir o influir de otra manera en la conducta que se observa.

Operación. En el modelo de la estructura del intelecto de J. P. Guilford, uno de los cinco tipos posibles de procesos mentales (cognoscitivo, memoria, pensamiento divergente, pensamiento convergente, evaluación). En la teoría del desarrollo cognoscitivo de J. Piaget, una operación es cualquier acción mental que es reversible (puede regresar a su punto de partida) e integrada con otras acciones mentales reversibles.

Operaciones formales. La etapa final (de los 11 a los 15 años) en la secuencia de desarrollo cognoscitivo de J. Piaget, en la cual el niño puede usar la lógica y el razonamiento verbal y realizar operaciones mentales más abstractas de nivel superior.

Opinión. Juicio verbalizado concerniente a una ocurrencia o situación específica. El significado de *opinión* es similar al de *actitud*, pero el primer término tiene la connotación de ser más específico y de basarse en más pensamiento que el último término. Además, una persona es consciente de sus opiniones, pero no necesariamente tiene conciencia de sus actitudes.

Orientación personal. Disposición de personalidad generalizada, como los roles de género o la autorrealización, que dirige la conducta en una variedad de situaciones.

Parkinsonismo. Trastorno cerebral progresivo que resulta del daño en los ganglios basales y ocurre más a menudo en la vejez. Los síntomas son temblores musculares, movimientos espásticos y rígidos; modo de andar propulsivo y rostro sin expresión, como una máscara. También se llama *enfermedad de Parkinson*.

Pensamiento convergente. Uso de hechos y de la razón para producir una sola respuesta a un problema.

Pensamiento divergente. Pensamiento creativo que implica más de una solución a un problema.

Percentil. La calificación de la prueba en la cual o por debajo de la cual cae un porcentaje especificado de calificaciones.

Perfil. Representación gráfica de la calificación de una persona en una serie de pruebas o subpruebas que componen un test compuesto o batería.

Personalidad. La suma total de todas las cualidades, rasgos y conductas que caracterizan a una persona y por la cual, junto con sus atributos físicos, la persona es percibida como un individuo.

Planteamiento ideográfico. Aproximación a la evaluación de la personalidad y la investigación en la cual se considera al individuo como un sistema legal e integrado por derecho propio. (Vea *planteamiento nomotético.*)

Planteamiento nomotético. Búsqueda de leyes generales de comportamiento que se apliquen a todos los individuos. (Vea *planteamiento ideográfico.*)

Población objetivo. La población de interés en la estandarización de una prueba u otro instrumento de evaluación; el grupo (muestra) de norma debe ser representativo de la población objetivo para poder hacer interpretaciones válidas de las calificaciones (referidas a la norma)

Ponderación de confianza. Procedimiento objetivo de calificación en el cual los pesos numéricos asignados a las respuestas correctas a los reactivos de la prueba dependen del grado de confianza declarado por el examinado de que sus respuestas son correctas.

Portafolio. Colección de los productos o desempeños de un estudiante a lo largo de un periodo que puede ser evaluada.

Potencial de aprendizaje estimado (ELP). Estimado de la capacidad de un niño para aprender, derivado de medidas obtenidas en el *Sistema de Evaluación Pluralista Multicultural* (SOMPA). El ELP toma en consideración no sólo el CI del niño en la Escala de Inteligencia para Niños de Wechsler-Revisada o la Escala de Inteligencia para Niveles Preescolar y Primaria de Wechsler, sino también el tamaño de la familia, la estructura familiar, el nivel socioeconómico y el grado de aculturación urbana.

Pregunta de alternativa fija. Pregunta de opción múltiple que consta de un tronco y varias respuestas posibles.

Preparación. Instrucción a corto plazo diseñada para mejorar las calificaciones de la prueba de los sujetos en prospectiva. Las actividades instruccionales incluyen práctica en varios tipos de reactivos y estrategias para presentar pruebas.

Procesamiento secuencial. Proceso mental en el cual una serie de reactivos se procesa secuencialmente en orden serial. Un ejemplo de una tarea secuencial es intentar recordar una serie de números. (Vea *procesamiento simultáneo.*)

Procesamiento simultáneo. Proceso mental en el cual varias piezas de información son sintetizadas o integradas de manera simultánea. (Vea *procesamiento secuencial.*)

Procesos componentes. De acuerdo con la teoría de Sternberg, los procesos cognoscitivos o componentes mentales, incluyendo metacomponentes, componentes de desempeño, componentes de adquisición, componentes de retención y componentes de transferencia.

Profecía que se cumple por sí misma. Tendencia a que las expectativas y actitudes de una persona concernientes a acontecimientos o resultados futuros afecten su ocurrencia; la tendencia de los niños a comportarse de las maneras que sus padres o maestros esperan que lo hagan.

Promedio. Medida de tendencia central de un grupo de calificaciones; la calificación más representativa.

Prueba adaptada. Procedimiento de prueba, por lo general basada en la computadora, en el cual los reactivos específicos que se presentan varían según la capacidad estimada u otras características especificadas del examinado y sus respuestas a los reactivos anteriores.

Prueba adaptada. (Vea *pruebas secuenciales.*)

Prueba con referencia a dominio. (Vea *prueba con referencia a criterio.*)

Prueba con referencia a criterio. Prueba diseñada con especificaciones de contenido muy restringidas para cumplir un rango limitado de propósitos muy específicos. El propósito de la prueba es determinar dónde se localiza el sujeto con respecto a ciertos objetivos educativos. (Vea *prueba con referencia a norma.*)

Prueba con referencia a normas. Prueba cuyas calificaciones son interpretadas con respecto a las normas obtenidas de una muestra representativa de individuos. (Vea *prueba con referencia.*)

Prueba de analogías. Prueba que requiere que el examinado determine una relación, similitud o diferencia entre dos o más cosas. Por ejemplo: "Las rosas son rojas como las violetas son (a) azules, (b) verdes, (c) naranjas, (d) amarillas".

Prueba de aptitudes académicas. Cualquier prueba que predice la habilidad de una persona para aprender los tipos de información y habilidades que se enseñan en la escuela. Las habilidades medidas por esas pruebas (por ejemplo, la prueba de aptitud académica) son similares a las medidas por las pruebas generales de inteligencia.

Prueba de aptitudes. Medida de la habilidad para beneficiarse del entrenamiento o experiencia adicionales, es decir, volverse competente en una destreza u otra habilidad.

Prueba de creatividad. Prueba que evalúa el pensamiento original, novedoso o divergente.

Prueba de diagnóstico. Prueba de aprovechamiento compuesta por diversas áreas o habilidades que constituyen cierta materia, con el propósito de diagnosticar las fortalezas y debilidades relativas de un individuo en esas áreas. Se dispone de pruebas de diagnóstico en lectura, aritmética y ortografía.

Prueba de ensayo. Prueba en la cual se requiere que los sujetos elaboren respuestas más bien extensas a una serie de preguntas. Las respuestas son evaluadas subjetivamente por el maestro u otro evaluador. (Vea *prueba objetiva.*)

Prueba de habilidades básicas. Medición de la competencia en lectura, matemáticas elementales u otras habilidades requeridas en la mayoría de los escenarios de entrenamiento y empleo.

Prueba de habilidades. Prueba que mide el grado en que una persona es capaz de realizar cierta tarea u ocupación.

Prueba de nivel flexible. Una prueba que consta de reactivos arreglados en orden de dificultad y en la cual cada examinado comienza por la mitad. Cuando un reactivo se responde correctamente, se presenta el siguiente

reactivo más difícil; cuando un reactivo se responde de manera incorrecta, se presenta el siguiente reactivo más fácil, y así sucesivamente.

Prueba de pico. Prueba diseñada para medir con eficiencia dentro de un rango bastante estrecho de habilidad.

Prueba de preparación. Prueba que mide el grado en que una persona posee las habilidades y el conocimiento necesarios para aprender una materia compleja como la lectura, las matemáticas o la ortografía.

Prueba de pronóstico. Prueba utilizada para predecir el aprovechamiento de una persona en una materia particular, por ejemplo, una prueba de preparación para la lectura.

Prueba de reconocimiento. En contraste con una prueba de diagnóstico, una prueba de aprovechamiento que se concentra en el desempeño global del examinado en la prueba.

Prueba de referencia. Grupo común de reactivos en cada una de las distintas formas de una prueba que se utiliza para igualar las calificaciones en las diversas formas.

Prueba de situación. Prueba de desempeño en la cual se coloca a la persona en una situación realista pero artificial y se le pide que cumpla una tarea específica. Las pruebas situacionales han sido empleadas para evaluar características de personalidad como la honestidad y la tolerancia a la frustración.

Prueba de una muestra de trabajo. Prueba que consiste en réplicas en miniatura de las tareas realizadas en el trabajo solicitado por el individuo. (Vea *test de réplica del trabajo.*)

Prueba justa para las culturas. Prueba compuesta por materiales a los cuales se supone que todos los grupos socioculturales han sido expuestos. La prueba no penaliza a ningún grupo sociocultural por la carencia de experiencia relevante. Los intentos por desarrollar pruebas justas para las culturas no han tenido mucho éxito.

Prueba objetiva. Prueba que se califica comparando las respuestas del sujeto con una lista de respuestas correctas (una clave) preparada de antemano, en contraste con una prueba que se califica de manera subjetiva. Algunos ejemplos de reactivos de las pruebas objetivas son los de opción múltiple y de verdadero-falso.

Prueba piloto. Prueba aplicada a una muestra representativa de personas para poner a prueba algunos aspectos de la prueba o de los reactivos de la prueba.

Prueba segura. Prueba aplicada bajo condiciones de alta seguridad para asegurarse de que sólo presenten la prueba las personas que deben hacerlo y que los examinados no saquen de la sala de exámenes copias de los materiales de prueba.

Pruebas en las que hay mucho en juego. Pruebas que contribuyen a tomar decisiones educativas, de empleo, de tratamiento u otras decisiones importantes concernientes a individuos o grupos.

Pruebas secuenciales. Procedimiento de examinación en el cual las respuestas del examinado a los reactivos previos determinan qué reactivos se presentarán a continuación; también se conoce como *pruebas adaptadas*.

Psicometría. Teoría e investigación concerniente a la medición de las características psicológicas (cognoscitivas y afectivas).

Pupilometría. Procedimiento para medir el diámetro de la pupila como un indicador del placer o interés en un estímulo específico.

***r*.** Símbolo para el coeficiente de correlación producto-momento de Pearson.

Rango percentilar. El porcentaje de calificaciones que caen por debajo de una calificación determinada en una distribución de frecuencia o grupo de calificaciones; el porcentaje correspondiente a la calificación determinada.

Rango semi-intercuartilar (*Q*). Medida de la variabilidad de un grupo de calificaciones de escala ordinal, calculadas como la mitad de la diferencia entre el primer y el tercer cuartil.

Rango. Medida cruda de la extensión o variabilidad de un grupo de calificaciones que se calcula restando la menor calificación de la calificación más alta.

Rapport. Relación cálida y amigable entre el examinador y el sujeto.

Razón CI. Cociente de inteligencia obtenido al dividir la calificación de edad mental del examinado en una prueba de inteligencia (como la antigua Stanford-Binet) entre su edad cronológica y multiplicando el cociente por 100. (Vea *CI de desviación*).

Razón de selección. La proporción de solicitantes que son seleccionados para un trabajo o programa de entrenamiento (educativo).

Reactivo abierto-cerrado. (Vea *reactivo de respuesta construida*.)

Reactivo de aparejamiento. Reactivo de una prueba que requiere que los individuos indiquen cuál(es) de varias opciones en una lista es (son) la(s) comparación (comparaciones) o respuesta(s) correcta(s) para cada una de las diversas opciones en otra lista.

Reactivo de elección forzada. Reactivo en un inventario de personalidad o de intereses, que se arregla en forma de díadas (dos opciones), tríadas (tres opciones) o tétradas (cuatro opciones). Se requiere que el individuo seleccione la opción que considere la más descriptiva de la personalidad, intereses o comportamiento de la persona que está siendo evaluada, y quizá otra opción percibida como la menos descriptiva de la personalidad, intereses o conducta de la persona evaluada. Los reactivos de

elección forzada se encuentran en ciertos inventarios de personalidad (por ejemplo, la Lista de Preferencias Personales de Edwards), inventarios de interés (el Estudio de Intereses Generales de Kuder) y formas de calificación para controlar los grupos de respuestas.

Reactivo de opción múltiple. Reactivo de una prueba que consta de un tronco (afirmación, pregunta, frase o similar) y varias opciones de respuesta (por lo general de tres a cinco), sólo una de las cuales es correcta.

Reactivo de reordenamiento. Reactivo de prueba en la cual se requiere que el examinado reordene los reactivos en su secuencia correcta.

Reactivo de respuesta construida. Pregunta o problema que se responde con una construcción escrita, pictórica, gráfica, etc. También se conoce como *reactivo abierto-cerrado.*

Reactivo de respuesta corta. Reactivo de prueba que requiere que el examinado construya una respuesta corta para llenar un espacio en blanco o para responder a una pregunta.

Reactivo de verdadero y falso. Reactivo de una prueba que consiste en una afirmación que es verdadera o falsa.

Reactivo. Una de las unidades, preguntas o tareas de las cuales está compuesto un instrumento psicométrico.

Reactivos entrelazados. Reactivos de una prueba en los cuales la respuesta a un reactivo es afectada o depende de las respuestas a otros reactivos de la prueba.

Registro anecdótico. Registro escrito de observaciones conductuales de un individuo especificado. Para que el registro sea objetivo debe tenerse cuidado en diferenciar entre observación e interpretación.

Regla de los cuatro quintos. Regla de selección que afirma que cualquier procedimiento que dé por resultado una tasa de selección para cualquier raza, género o grupo étnico que sea menor de cuatro quintos (80%) de la del grupo con la tasa más alta tiene un efecto adverso y en consecuencia es ilegal.

Regla de oro. Compromiso establecido entre la compañía de seguros Golden Rule y el Servicio de Pruebas Educativas, en el cual este último aceptaba elaborar un examen para los agentes de seguros que constara de reactivos que mostraran la menor cantidad de discrepancia entre los solicitantes blancos y negros.

Regresión hacia la media. Tendencia de las calificaciones de una prueba o de otras medidas psicométricas a acercarse a la media al volver a aplicar la prueba; entre más extrema sea la calificación original, más cerca estará de la media en la segunda prueba.

Rendimiento. Grado de éxito o logro en un área o empeño dado; una puntuación en una prueba de rendimiento.

Retrasado mental susceptible de ser capacitado (TMR). Niños en el rango de CI de retraso moderado (aproximadamente de 36 a 50), quienes por lo general no pueden aprender a leer y escribir pero pueden realizar bajo supervisión tareas que no requieren de gran habilidad.

Retrasado mental susceptible de ser educado (RME). Niños que se caracterizan por un grado leve de retraso mental (CI de 51 a 69). Dichos niños son capaces de obtener una educación de tercer a sexto grados y pueden aprender a leer, escribir y a realizar operaciones aritméticas elementales.

Retrasado mental. Una persona cuyo funcionamiento intelectual está significativamente por debajo del promedio, por lo general se define como un CI de 70 o 75 y más bajo.

Rotación de factores. Procedimiento matemático aplicado a una matriz de factores con el propósito de simplificar la matriz para su interpretación incrementando el número de cargas de factores altas y bajas en la matriz. La rotación de factores puede ser *ortogonal*, en cuyo caso los factores resultantes forman ángulos rectos entre sí, u *oblicua*, en la cual los ejes del factor resultante forman ángulos agudos u obtusos entre sí.

Rotación oblicua. En el análisis factorial, una rotación en la cual se permite que los ejes del factor formen ángulos agudos u obtusos entre sí. En consecuencia, los factores están correlacionados.

Rotación ortogonal. En el análisis factorial, una rotación que mantiene la independencia de los factores; es decir, los ángulos entre los factores se mantienen a 90 grados y por ende los factores no están correlacionados.

Rúbrica. Procedimiento, esquema o conjunto de criterios de acuerdo con los cuales se evalúa el desempeño.

Sagacidad para la prueba. Técnicas diferentes al conocimiento del material de la prueba que utilizan los examinados para mejorar sus calificaciones en la prueba.

Selección. El uso de pruebas y otros instrumentos para seleccionar a aquellos solicitantes de un trabajo o un programa educativo que tengan más probabilidad de tener éxito en esa situación. Los solicitantes que obtienen la calificación límite o más arriba son seleccionados (aceptados) y los que caen por debajo del límite son rechazados.

Sesgo. Cualquiera de una serie de factores que ocasiona que las calificaciones en los instrumentos psicométricos sean consistentemente mayores o menores de lo que serían si la medición fuera exacta. Un ejemplo de los factores que resultan en sesgo es el *error de indulgencia*, la tendencia a calificar a una persona consistentemente más alto de lo que debería ser.

Sesgo de confirmación. Tendencia a buscar y recordar información que es congruente con las creencias o preconcepciones propias.

Sesgo de creencia. La tendencia a que las creencias ya existentes distorsionen el razonamiento lógico, hacien-

do que las conclusiones no válidas parezcan válidas o que las conclusiones válidas parezcan no válidas.

Sesgo de la prueba. Condición en la cual una prueba discrimina de manera injusta entre dos o más grupos.

Sesgo del reactivo. Grado en el cual un reactivo mide constructos diferentes en distintos grupos étnicos, culturales, regionales o de género.

Sesgo. Grado de asimetría en una distribución de frecuencia. En una distribución con sesgo positivo hay más calificaciones a la izquierda de la moda (bajas calificaciones); esto es cierto cuando la prueba es demasiado difícil para los examinados. En una distribución con sesgo negativo hay más calificaciones a la derecha de la moda (altas calificaciones); esto es cierto cuando la prueba es demasiado fácil para los examinados.

Síndrome de Down (mongolismo). Trastorno caracterizado por un cráneo aplanado, piel engrosada en los párpados, dedos cortos y gruesos; cabello grueso y sedoso; pequeña estatura e inteligencia moderadamente baja. Se encuentra un cromosoma extra en la posición veintiuno del cariotipo de los casos del síndrome de Down.

Sociograma. Diagrama que consta de círculos que representan a los individuos en un grupo, con líneas que se trazan para indicar qué personas se eligen (aceptan) entre sí y qué personas no se eligen (rechazan) entre sí. Los términos usados al referirse a los elementos particulares de un sociograma son *estrella, pandilla, aislado* y *sociedad de admiración mutua.*

Somatotipo. Clasificación de la estructura corporal (físico) en el sistema de tres componentes de Sheldon (endomorfia, mesomorfia, ectomorfia).

Sublimación. Desviación de la energía de un impulso sexual u otro impulso biológico de su meta inmediata a una naturaleza o uso social, moral o estética más alta.

Subtest. Parte o subgrupo de reactivos de una prueba (por ejemplo, un grupo de reactivos que miden la misma función o reactivos en el mismo nivel de edad o nivel de dificultad).

Superdotado. Persona cuyo funcionamiento intelectual está significativamente por arriba del promedio, por lo general se define como un CI de 130 o 140 y más alto.

Tabla de especificaciones. Tabla con dos entradas preparada como un bosquejo o esqueleto de una prueba de aprovechamiento. Los objetivos conductuales se encuentran en los encabezados de las hileras y los objetivos de contenido (temas) en los encabezados de las columnas de dicha tabla.

Tabla de expectativas. Tabla que ofrece la frecuencia o porcentaje de examinados en una cierta categoría (intervalo de calificaciones) en una variable de predicción (prueba) que se espera caiga en cierta categoría (intervalo de calificaciones) en la variable criterio.

Tabla de Snellen. Tabla que contiene letras de varios tamaños, diseñada para medir la agudeza visual a cierta distancia.

Tablas de Taylor-Russell. Tablas para evaluar la validez de una prueba como función de la información aportada por la prueba más allá de la información aportada por el azar.

Tasa base. Proporción de individuos en una población especificada que posee cierta característica, condición o conducta.

Taxonomía de conducta psicomotriz. Taxonomía de objetivos para una lección o prueba en un dominio psicomotriz.

Taxonomía de objetivos educativos. Conjunto arreglado jerárquicamente, mutuamente inclusivo de objetivos para una lección o prueba de aprovechamiento.

Técnica Cloze. Procedimiento de prueba en el cual se borran palabras al azar de un pasaje escrito y se pide al examinado que las reemplace. El grado en el que el examinado puede dar sentido al pasaje y en que logra llenar los espacios en blanco es una medida de su habilidad de lectura.

Técnica de "adivina quién". Procedimiento para analizar la interacción de grupo y el valor del estímulo social de los miembros del grupo, en el cual se pide a los niños que "adivinen quién" posee ciertas características o hace ciertas cosas en un aula u otra situación de grupo.

Técnica de la charola de pendientes. Procedimiento para la evaluación de supervisores o ejecutivos en la cual se pide al candidato que indique qué acción debería tomarse sobre una serie de memoranda y otros materiales del tipo que suele encontrarse en la charola de pendientes de un supervisor o ejecutivo.

Técnica de nominación. Método para estudiar la estructura social y la personalidad en el cual se pide a los estudiantes, trabajadores u otros grupos de individuos que indiquen con qué personas del grupo les gustaría hacer cierta cosa o quiénes creen que poseen ciertas características.

Técnica proyectiva. Técnica de evaluación de personalidad relativamente no estructurada en la cual la persona responde a materiales como manchas de tinta, ilustraciones ambiguas, frases incompletas y otros materiales diciendo lo que percibe, elaborando historias o construyendo y arreglando frases y objetos. En teoría, como el material no es estructurado, la estructura impuesta sobre éste por las respuestas del examinado representa una proyección de sus propias características de personalidad (necesidades, conflictos, fuentes de ansiedad, etcétera.)

Técnica *Q* (clasificación *Q*). Procedimiento de evaluación de la personalidad que se concentra en la clasificación de tarjetas (clasificaciones *Q*) que contienen afirmaciones que pueden ser o no descriptivas del calificador.

Técnica sociométrica. Método para determinar y describir el patrón de aceptaciones y rechazos en un grupo de personas.

Tendencia central. Calificación promedio, o central, en un grupo de calificaciones; la calificación más representativa (por ejemplo, media aritmética, mediana, moda).

Teoría de la generalización. Teoría de las calificaciones de la prueba y la formulación estadística asociada que conceptualiza una calificación de la prueba como una muestra de un universo de calificaciones. Se utilizan procedimientos de análisis de varianza para determinar la generalización de la calificación al valor universal, como función de los examinados, los reactivos de la prueba y los contextos situacionales. Puede calcularse un coeficiente de generalización como una medida del grado de generalización de la muestra a la población.

Teoría de rasgos latentes. Cualquiera de varias teorías (por ejemplo, la teoría de la curva característica de los reactivos, el modelo Rasch) y los procedimientos estadísticos asociados que relacionan las calificaciones en el reactivo y la prueba con la posición estimada en algún rasgo de habilidad latente hipotético o continuo; se utiliza en el análisis de reactivos y en la estandarización de la prueba.

Teoría de respuesta a los ítemes (IRT). Teoría de los reactivos de una prueba en el cual las calificaciones del reactivo se expresan en términos de calificaciones estimadas en un continuo de habilidad latente.

Teoría del espejo. De acuerdo con C. H. Cooley, la idea de que el yo se forma como resultado de la percepción que tiene el individuo de cómo ven los otros a su persona y su conducta.

Teratógeno. Sustancia (alcohol, drogas, etc.) que cruza la barrera placentaria entre la madre y el feto y ocasiona deformidad física en el feto.

Test. Cualquier instrumento utilizado para evaluar la conducta o el desempeño de una persona. Las pruebas psicológicas son de muchas clases: cognoscitivas, afectivas y psicomotrices, por ejemplo.

Test colectivo en espiral. Prueba que consiste en una variedad de reactivos dispuestos en orden de dificultad creciente. Los reactivos de un tipo dado de contenido aparecen en toda la prueba entremezclados con otros tipos de reactivos de dificultad similar en una espiral de dificultad creciente.

Test colectivo. Una prueba que consta de una variedad de reactivos diseñados para medir diferentes aspectos del funcionamiento mental. El Test de Habilidad Escolar de Otis-Lennon y el Test Henmon-Nelson de Habilidad Mental son pruebas colectivas. (Vea *test colectivo en espiral.*)

Test comercial. Prueba de conocimiento y habilidad en una ocupación particular; se usa para la selección de personal, colocación y otorgamiento de licencia.

Test de asociación de palabras. Prueba proyectiva en la cual el examinado responde a cada una de las palabras presentadas por el examinador con la primera palabra que le venga a la mente. Las respuestas inusuales o el responder de manera lenta a ciertas palabras pueden ser indicativos de conflictos u otros problemas emocionales asociados con esas palabras.

Test de dominio. (Vea *prueba con referencias a criterios.*)

Test de ejecución. Prueba en la que se requiere que el individuo manipule diversos objetos físicos; las pruebas de ejecución contrastan con las de lápiz y papel. Algunos ejemplos son la escala de desempeño de la escala de inteligencia de Wechsler y la Escala de Desempeño de Arthur Point.

Test de frases incompletas. Prueba proyectiva de personalidad que consta de una serie de frases incompletas que el examinado debe completar.

Test de grupo. Prueba que un examinador aplica de manera simultánea a un grupo de individuos. (Vea *test individual.*)

Test de inteligencia. Prueba psicológica diseñada para medir la aptitud de un individuo para el trabajo escolar o para otra clase de actividades que impliquen habilidad verbal y solución de problemas.

Test de lenguaje. Prueba compuesta por reactivos verbales o numéricos, es decir, reactivos que implican el uso del lenguaje. (Vea *test no verbal.*)

Test de niveles múltiples. Prueba diseñada para ser apropiada para varios niveles de edad; se elabora una prueba separada para cada nivel.

Test de poder. Una prueba con límite de tiempo amplio, de modo que todos los examinados tienen tiempo para tratar de responder a todos los reactivos. Muchos de los reactivos son difíciles, y a menudo se presentan en orden de dificultad del más fácil al más difícil.

Test de réplica del trabajo. Una prueba en la que se requiere que el individuo realice un conjunto de operaciones o tareas similares a las del trabajo real. También se conoce como *test de muestra de trabajo.*

Test de velocidad. Prueba que consiste en un gran número de reactivos sencillos, pero que tiene un tiempo límite corto por lo que casi nadie completa la prueba en el tiempo permitido. Muchas pruebas de habilidad para el trabajo de oficina, mecánica y psicomotriz son de velocidad.

Test estandarizado. Prueba que ha sido elaborada cuidadosamente por profesionales y aplicada con direcciones estándar y bajo condiciones estándar. La prueba se apli-

ca usualmente a una muestra representativa de personas con el propósito de obtener normas.

Test fuera de nivel. Aplicación de una prueba diseñada principalmente para una edad o nivel a individuos que se encuentran por debajo o por arriba de ese nivel.

Test individual. Prueba que se aplica a una persona a la vez.

Test no verbal. Prueba que no necesita el uso de palabras habladas o escritas, sino que requiere que el individuo construya, manipule o responda a los materiales de la prueba de otras formas no verbales.

Test oral. Prueba en que el individuo proporciona respuestas orales a preguntas orales o escritas.

Test verbal. Prueba con instrucciones verbales que requieren respuestas orales o escritas con palabras y/o números.

Traslape de reactivos. Grado en el cual dos reactivos miden la misma variable y, en consecuencia, las calificaciones en los reactivos están correlacionadas.

Trastorno neuropsicológico. Trastorno del sistema nervioso acompañado por síntomas psicológicos.

Ubicación en grado por edad mental. Indicador del nivel de grado al que corresponde el funcionamiento mental de una persona.

Validación cruzada. Aplicar de nuevo un instrumento de evaluación que se considera un predictor válido de un criterio para un grupo de personas a un segundo grupo de personas para determinar si el instrumento también es válido para ese grupo. En la validación cruzada casi siempre existe una reducción del coeficiente de validez, ya que los factores del azar aumentan de manera espuria el coeficiente de validez obtenido con el primer grupo de examinados.

Validez. El grado en el que un instrumento de evaluación mide lo que está diseñado para medir. La validez puede ser evaluada de varias maneras: mediante el análisis del contenido del instrumento (*validez de contenido*), relacionando las calificaciones en la prueba con un criterio (*validez de predicción y concurrente*) y mediante un estudio más profundo del grado en que la prueba es una medida de cierto constructo psicológico (*validez de constructo*).

Validez aparente. El grado en que la apariencia o contenido del material (reactivos y similares) de una prueba u otro instrumento psicométrico es tal que el instrumento parece ser una buena medida de lo que debe medir.

Validez con relación a criterios. El grado en que una prueba u otro instrumento de evaluación miden lo que están diseñados para medir, según lo indica la correlación de las calificaciones de la prueba con alguna medida criterio del comportamiento.

Validez concurrente. El grado en el que las calificaciones obtenidas por un grupo de personas en un instrumento psicométrico particular se relacionan con sus calificaciones determinadas simultáneamente en otra medida (criterio) de las mismas características que el instrumento pretende medir.

Validez convergente. Situación en la cual un instrumento de evaluación tiene correlaciones elevadas con otras medidas (o métodos de medición) del mismo constructo. (Vea *validez discriminante*.)

Validez creciente. Un incremento en la validez producido por una nueva prueba por encima de la obtenida con los procedimientos existentes de selección.

Validez de constructo. El grado en el cual las calificaciones de un instrumento psicométrico diseñado para medir ciertas características se relacionan con las medidas del comportamiento en situaciones en las que se supone que la característica es un determinante importante del comportamiento.

Validez de contenido. El grado en que un grupo de expertos en el material del que trata una prueba están de acuerdo en que la prueba u otro instrumento psicométrico miden lo que están diseñados para medir.

Validez discriminante. Situación en la cual un instrumento psicométrico tiene correlaciones bajas con otras medidas (o métodos de medición) de diferentes constructos psicológicos.

Validez predictiva. Grado en el cual las calificaciones en una prueba pueden predecir el desempeño en alguna medida criterio en un momento posterior; por lo general se expresa como una correlación entre la prueba (variable predictora) y la variable criterio.

Variabilidad. El grado de dispersión o desviación de un grupo de calificaciones alrededor de su valor promedio.

Variable moderadora. Variable demográfica o de personalidad (por ejemplo, edad, sexo, estilo cognoscitivo, compulsividad) que afecta la correlación entre otras dos variables (por ejemplo, aptitud y aprovechamiento).

Variable. En contraste con una *constante*, cualquier cantidad que pueda asumir más de un estado o valor numérico.

Varianza. Medida de variabilidad de las calificaciones de una prueba; se calcula como la suma de los cuadrados de las desviaciones de las calificaciones crudas respecto a la media aritmética, divididas entre el número de calificaciones menos uno; el cuadrado de la desviación estándar.

Yo ideal. En la teoría fenomenológica de C. R. Rogers, la persona que le gustaría ser al individuo, en contraste con el *yo real* de la persona.

Zona de desarrollo potencial. La diferencia (distancia) entre el nivel de desarrollo actual de un niño (su desarrollo completado tal como puede ser evaluado por una prueba estandarizada) y su potencial de desarrollo (el grado de competencia que puede obtener con ayuda).

RESPUESTAS A LAS ACTIVIDADES Y PREGUNTAS CUANTITATIVAS

CAPÍTULO 3

6. Calificación no corregida = número correcto = 30.

Calificación corregida = aciertos – equivocaciones/3 = 30 – 16/3 ≈ 25.

Si los reactivos son verdadero-falso, calificación no corregida = número correcto = 30, y calificación corregida = aciertos – equivocaciones = 14.

7. La suma de los valores absolutos de las diferencias entre las clasificaciones con clave y la clasificación de Juan es 12. Usando la fórmula 3.1a, su calificación es 3.5, la cual se redondea a 4. En el caso de María, la suma de los valores absolutos de las diferencias es 6, y su calificación se redondea a 5. Usando la fórmula 3.1b, las calificaciones redondeadas para Juan y María son 4 y 6, respectivamente.

8. Prueba *X*:

GRADO	RANGO	NÚMERO
A	44 y más	2
B	38–43	7
C	31–37	12
D	25–30	6
F	24 y menos	3

Prueba *Y*:

GRADO	RANGO	NÚMERO
A	44 y más	3
B	33–43	5
C	22–32	13
D	11–21	7
F	10 y menos	2

CAPÍTULO 4

1. Ya que .27 × 75 = 20.25, hay 20 personas en el grupo superior y 20 en el grupo inferior. Por lo tanto, $p = (18 + 12)/40 = .75$ y $D = (18 – 12)/20 = .30$. El reactivo está en los rangos aceptables tanto de p como de D.

2. $U = 30$, $L = 20$, $U_p = 20$, y $L_p = 10$, de modo que $p = (20 + 10)/50 = .60$ y $D = 20/30 – 10/20 = .17$.

3.

REACTIVO	1	2	3	4	5	6	7	8	9	10
p	.50	.45	.45	.55	.40	.75	.50	.50	.60	.40
D	.40	.30	.30	.50	.60	.30	.20	.20	.40	.60

La tabla 4.1 da el valor promedio de *p* óptimo de un reactivo de opción múltiple de cuatro opciones como .74. Tomando ±.20 alrededor de este valor, los reactivos aceptables deberían estar en el rango *p* de .54 a .94. El valor *D* de los reactivos aceptables debería ser de .30 o mayor. De acuerdo con estos criterios, sólo los reactivos 4, 6 y 9 son aceptables. Los restantes siete reactivos deberían ser modificados o eliminados.

4. La calificación *z* de Jorge en la prueba de aritmética es $z_a = (65 – 50)/10 = 1.50$; su calificación *z* en la prueba de lectura es $z_r = (80 – 75)/15 = .33$. Sus calificaciones *Z* en las dos pruebas son $Z_a = 10 (1.5) + 50 = 65$ y $Z_r = 10 (.3) + 50 = 53$. Por lo tanto, Jorge está ligeramente mejor en aritmética que en lectura.

5.

% RANGO	*z*	*T*	CEEB	ESTANINA	DESVIACIÓN CI
10	–1.28	37	372	2	81
20	–.84	42	416	3	87
30	–.52	45	448	4	92
40	–.25	48	475	4	96
50	.00	50	500	5	100
60	.25	52	525	6	104
70	.52	55	552	6	108
80	.84	58	584	7	113
90	1.28	63	628	8	119

6.

INTERVALO DE CALIFICACIÓN	PUNTO MEDIO	FRECUENCIA	FRECUENCIA ACUMULATIVA BAJO PUNTO MEDIO	*Z*	z_n	*Z*
69–71	97	1	70	2.13	71
96–98	94	2	65	1.50	65
93–95	91	3	29.5	61	1.04	60
90–92	88	5	28	56	.57	56
87–89	85	5	25.5	52	.13	51
84–86	82	5	21.5	47	–.30	47
81–83	79	4		43	–.73	43
78–80	76	2		38	–1.11	39
75–77	73	2		33	–1.50	35
72–74	70	1		29	–2.13	29

CAPÍTULO 5

1. $r_{oe} = .226$, $r_{11} = .369$, $KR_{20} = .610$, $KR_{21} = .580$.

2. $s_{err} = 4.00$

95% del intervalo de confianza para $X = 40$ es 32.16-47.84.

95% del intervalo de confianza para $X = 50$ es 42.16-57.84.

95% del intervalo de confianza para $X = 60$ es 52.16-67.84.

3. Sustituyendo en la fórmula 5.9, tenemos $m = .90\ (1 - .80)/[.80\ (1 - .90)] = .18/.08 = 2.25$. Multiplicando n por m da $40 \times 2.25 = 90$. Por lo tanto, 50 reactivos más del mismo tipo general que los de la prueba deben sumarse a la prueba para aumentar su coeficiente de confiabilidad a .90

5. $s_{est} = s\sqrt{1 - r^2} = .5\sqrt{1 - .60^2} = .5(8) = .4$. La probabilidad es .68 de que el promedio de puntos del grado del alumno caerá entre 2.1 y 2.9, y .95 de que caerá entre 1.72 y 3.28.

6.

VARIABLE PREDICTIVA	VARIABLE DE CRITERIO (Y)					
	26–34	*35–41*	*42–48*	*49–55*	*56–62*	*63–70*
71–78			1(100)			1(50)
63–70				1(100)		2(67)
56–62			2(100)	2(67)	1(33)	1(17)
49–55		2(100)	1(67)	2(50)		1(17)
42–48		1(100)	1(80)	2(60)		1(20)
35–41		2(100)	1(50)	1(25)		
28–34	1(100)			1(67)	1(33)	
21–27	1(100)					

CAPÍTULO 7

2. CI = 100(MA/CA) = 100(77/105) = ≈ 73.

CAPÍTULO 8

8. $h^2 = .65$, y significa que 65% de la varianza en las calificaciones CI es atribuible a factores genéticos.

CAPÍTULO 10

8. Sí. $s_{est} = 10\sqrt{2 - .90 - .85} = 5$, y $2 \times 5 = 10$ es igual a la diferencia entre las dos calificaciones *T*.

CAPÍTULO 11

7. La ecuación de regresión para predecir Y a partir de X es $Y_{pred} = .44X + 28.34$ para el grupo mayoritario, y $Y_{pred} = .43X + 24.57$ para el grupo minoritario; los coeficientes de correlación correspondientes son .52 y .47. El coeficiente de correlación sugiere que la prueba puede predecir ligeramente mejor para el grupo mayoritario que para el grupo minoritario, pero no de manera considerable. Así, puede concluirse que la prueba no está sesgada de forma perceptible de acuerdo con la definición tradicional de justicia.

Suponiendo que 50% del grupo mayoritario y 25% del grupo minoritario puede desempeñar el trabajo, entonces deberían seleccionarse .50(30) = 15 examinados X = 52, entonces se elegirán 7 miembros del grupo minoritario y 13 del grupo mayoritario. De acuerdo con la definición de Thorndike, la primera calificación límite favorecería ligeramente al grupo minoritario y la segunda desfavorecería un poco al grupo mayoritario.

Suponiendo que 40% de todo el grupo de 50 examinandos es capaz de desempeñar el trabajo, entonces .40(30) = 12 miembros del grupo mayoritario y .40(20) = 8 miembros del grupo minoritario deberían ser seleccionados, si la prueba es justa de acuerdo con la definición de Cole. Cualquier calificación límite que produzca un número total de seleccionados cercano a 20 tenderá a favorecer al grupo mayoritario según esta definición. Así, usando este procedimiento, la prueba está ligeramente sesgada en favor del grupo mayoritario.

Combinar las calificaciones de los grupos mayoritario y minoritario produce una correlación entre X y Y de r = .517 y la ecuación de regresión $Y_{pred} = .46X + 25.639$. Si la calificación límite se establece en X = 50, la cantidad de errores de falso positivo y falso negativo de cada grupo se distribuirá como sigue:

	FALSOS POSITIVOS	FALSOS NEGATIVOS
Grupo mayoritario	6 (20%)	6 (20%)
Grupo minoritario	5 (25%)	2 (10%)

Estos porcentajes están basados en la cantidad total de candidatos de cada grupo. El porcentaje de errores es 5 puntos mayor para el grupo mayoritario que para el minoritario, pero el porcentaje de falsos positivos es mayor en el grupo minoritario y el porcentaje de falsos negativos es superior en el grupo mayoritario. En este caso, el problema del sesgo es complejo, dependiendo de qué tipo de error se considere más grave.

CAPÍTULO 13

1.

ENUNCIADO	VALOR DE ESCALA (MEDIANA)	ÍNDICE DE AMBIGÜEDADES (Q)
D	8.96	1.14
N	5.19	1.22
X	2.50	1.01

4.

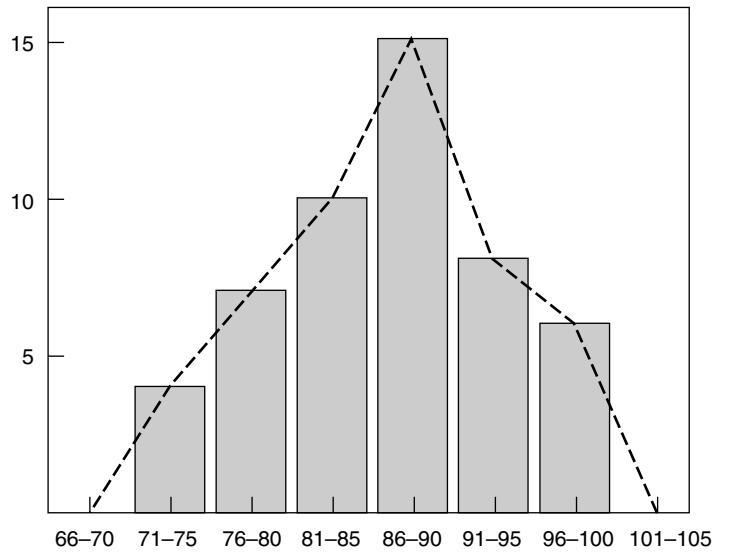
	Enunciado de Actitudes						
PARTICIPANTE	*1*	*2*	*3*	*4*	*5*	*6*	**TOTAL "+"**
A	+	+	+	+	+	+	6
D	+	+	+	+	+	–	5
E	–	–	+	+	+	+	4
B	–	–	+	+	+	–	3
F	–	–	+	–	+	+	3
G	–	–	+	+	–	–	2
C	–	–	–	–	+	–	1
Total "+"	2	2	6	5	6	3	
Errores	0	0	0	2	2	2	

$R = 1 - 6/(7 \times 6) = .857$

El coeficiente de reproductibilidad es inferior a .90 y, por ende, no constituye evidencia de que los seis enunciados formen una verdadera escala de Guttman.

APÉNDICE A

1.



$\bar{x} = 86.40$, Mdn = 86.83, Moda = 88, $s = 7.17$

$P_{25} = Q_1 = 80.50 + 5(12.5 - 11)/10 = 81.25$

$P_{75} = Q_3 = 90.5 + 5(37.5 - 36)/8 = 91.44$

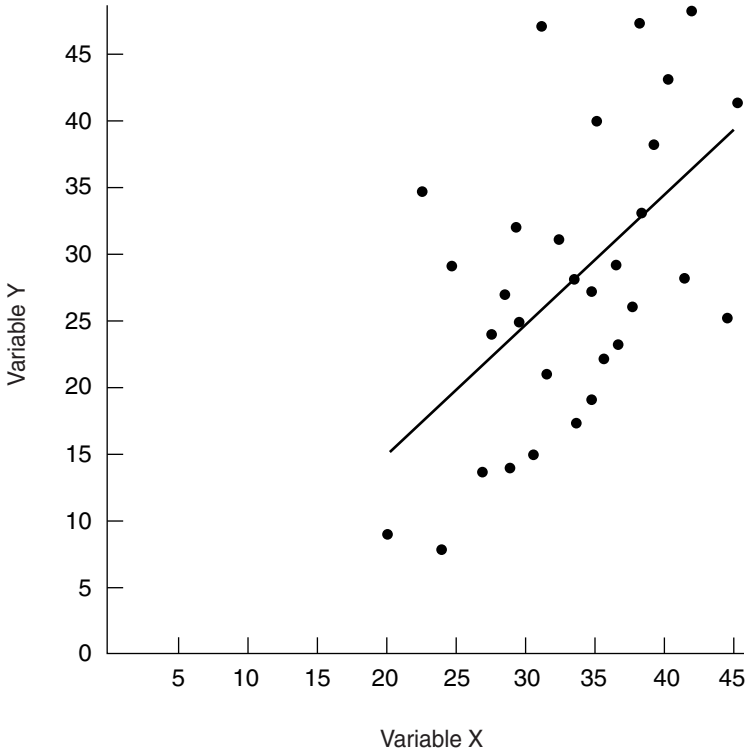
$Q = (Q_3 - Q_1)/2 = (91.44 - 81.25)/2 = 5.10$

2. ..5%, 1%, 2.5%, 5%, 50%, 95%, 97.5%, 99%, 99.5%
-1.28, -.84, -.52, -.25, .00, .25, .52, .84, 1.28

3. $\overline{x} = 34.00$, $s_x = 6.58$

$\overline{y} = 27.00$, $s_y = 11.03$

$r_{y \cdot x} = .54$, $Y_{\text{pred}} = .91X - 3.78$



REFERENCIAS

Abrahams, N. M., Neumann, I. y Gilthens, W. H. (1971). Faking vocational interests: Simulated vs. real life motivation. *Personnel Psychology, 24,* 5–12.

Achenbach, T. M. y Edelbrock, C. (1983). *Manual of the Child Behavior Checklist and Revised Child Behavior Profile.* Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M. y Edelbrock, C. (1986). *Manual for the Teacher's Report Form and Teacher Version of the Child Behavior Profile.* Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M. y Edelbrock, C. (1987). *Manual for the Youth Self-Report and Profile.* Burlington, VT: University of Vermont, Department of Psychiatry.

Aiken, L. R. (1970). Scoring for partial knowledge of the generalized rearrangement item. *Educational and Pyschological Measurement, 30,* 87–94.

Aiken, L. R. (1979). Attitudes toward mathematics and science in Iranian middle schools. *School Science and Mathematics, 79,* 229–234.

Aiken, L. R. (1980). Problems in testing the elderly. *Educational Gerontology, 5,* 119–124.

Aiken, L. R. (1983a). *The case for oral achievement testing.* ERIC Document Reproduction Service No. ED 222 578 & TM 820 755.

Aiken, L. R. (1983b). Determining grade boundaries on classroom tests. *Educational & Psychological Measurement, 3,* 759–762.

Aiken, L. R. (1983c). Number of response categories and statistics on a teacher rating scale. *Educational & Psychological Measurement, 43,* 397–401.

Aiken, L. R. (1988). KAPPO: A program for assessing the reliability of criterion-referenced tests. *Applied Psychological Measurement, 12,* 104.

Aiken, L. R. (1996). *Rating scales & checklists: Evaluating behavior, personality, and attitudes.* New York: Wiley.

Aiken, L. R. (1997). *Questionnaires & inventories: Surveying opinions and assessing personality.* New York: Wiley.

Aiken, L. R. (1998). *Tests & examinations: Measuring abilities and performance.* New York: Wiley.

Aiken, L. R. (1999). *Human differences.* Mahwah, NJ: Lawrence Erlbaum.

Aiken, L. R. (2000). Computer programs for facilitating objective grading. *Educational Research Quarterly, 24*(2), 55–61.

Airasian, P. W. y Terrasi, S. (1994). Test administration. En T. Husén & T. N. Postlethwaite (Eds.), *International encyclopedia of education* (2a. ed., Vol. 11, pp. 6311–6315). Tarrytown, NY: Elsevier.

Ajzen, I. y Fishbein, M. (1977). Attitude–behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin, 84,* 888–918.

Albemarle Paper Company v. *Moody.* 10 FEP 11 1181 (1975).

Albright, L. y Malloy, T. E. (1999). Self-observation of social behavior and metaperception. *Journal of Personality & Social Psychology, 77,* 726–734.

Alderton, D. L. (1994). Mechanical ability. En R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 697–700). New York: Macmillan.

Alhberg, J., Tuck, J. R. y Allgulander, C. (1996). Pilot study of the adjunct utility of a computer-assisted Diagnostic Interview Schedule (C-DIS) in forensic psychiatric patients. *Bulletin of the American Academy of Psychiatry & the Law, 24,* 109–116.

Allard, G., Butler, J., Faust, D. y Shea, M. T. (1995). Errors in hand scoring objective personality tests: The case of the Personality Diagnostic Questionnaire. *Professional Psychology: Research and Practice, 26,* 304–208.

Allard, G. y Faust, D. (2000). Errors in scoring objective personality tests. *Assessment, 7,* 119–129.

Allison, D. E. (1984). The effect of item-difficulty sequence, intelligence, and sex on test performance, reliability, and item difficulty and discrimination. *Measurement and Evaluation in Guidance, 16,* 211–217.

Allport, G. W. (1937). *Personality: A psychological interpretation.* New York: Holt, Rinehart & Winston.

Allport, G. W. (1961). *Pattern and growth in personality.* New York: Holt, Rinehart & Winston.

Allport, G. W. (Ed.). (1965). *Letters from Jenny.* New York: Harcourt Brace Jovanovich.

Allport, G. W. y Odbert, H. S. (1936). Trait-names. A psycholexical study. *Psychological Monographs, 47,* Bi, 211 161,

Allport, G. W., Vernon, P. E. y Lindzey, G. (1960). *Study of Values (3rd ed.): Manual.* Chicago: Riverside.

Altus, W. D. (1966). Birth order and its sequelae. *Science, 151,* 44–49.

Alwin, D. F. y Krosnick, J. A. (1991). The reliability of survey attitude measurement. *Sociological Methods & Research, 20,* 139–181.

American Association of Mental Retardation. (1992). *Mental retardation: Definition, classification, and systems of supports* (9a. ed.). Washington, DC: Author.

American College (1978). *Test wiseness: Test taking skills for adults.* New York: McGraw-Hill.

American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Law Institute (1956). *Model penal code.* Tentative Draft Number 4.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4a. ed.). Washington, DC: Author.

American Psychological Association. (1981). Ethical principles of psychologists. *American Psychologist, 36,* 633–638.

American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist, 47,* 1597–1611.

American Psychological Association, Committee on Professional Standards y Committee on Psychological Tests and Assessment. (1986). *Guidelines for computer-based tests and interpretations.* Washington, DC: American Psychological Association.

America's children: Key national indicators of well being. (1998). Washington, DC: Interagency Forum on Child and Family Statistics.

Ames, L. B. (1967). Predictive value of infant behavior examinations. En J. Hellmuth (Ed.), *Exceptional infant. Vol. 1: The normal infant* (pp. 207–239). Seattle: Straub & Hellmuth.

Ames, L. B., Gillespie, B. S., Haines, J. y Ilg, F. L. (1979). *The Gesell Institute's child from one to six: Evaluating the behavior of the preschool child.* New York: Harper & Row.

Anastasi, A. y Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Andreasen, N. C. (1987). Creativity and mental illness: Prevalence rates in writers and their first-degree relatives. *American Journal of Psychiatry, 144,* 1288–1297.

Anrig, G. R. (1987). "Golden Rule": Second thoughts. *APA Monitor, 18*(8), 3.

Ansley, T. (1997). The role of standardized achievement tests in grades K–12. En G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement, and adjustment* (pp. 265–285). San Diego, CA: Academic Press.

APA task force releases final report on integrity testing. (1991, mayo/junio). *Psychological Science Agenda 4*(3), pp. 1, 6. Washington, DC: American Psychological Association.

Archer, R. P., Maruish, M., Imhof, E. A. y Piotrowski, C. (1991). Psychological test usage with adolescent clients: 1990 survey findings. *Professional Psychology: Research and Practice, 22,* 247–252.

Arkes, H. R. (1994). Clinical judgment. En R. J. Corsini (Ed.), *Concise encyclopedia of psychology* (2a. ed., pp. 237–238). New York: Wiley.

Arvey, R. D. (1979). Unfair discrimination in the employment interview: Legal and psychological aspects. *Psychological Bulletin, 86,* 736–765.

Ash, P. (1995). Review of the Eating Disorder Inventory-2. *Twelfth Mental Measurements Yearbook,* 334–335.

Austin, G. R. y Garber, H. (Eds.). (1982). *The rise and fall of national test scores.* New York: Academic Press.

Baker, E. L., O'Neil, H. F. y Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48,* 1210–1218.

Baller, W. R., Charles, D. C. y Miller, E. L. (1967). Midlife attainment of the mentally retarded: A longitudinal study. *Genetic Psychology Monographs, 75,* 235–329.

Baltes, P. B. y Schaie, K. W. (1974). The myth of the twilight years. *Psychology Today, 7*(10), 35–40.

Baltes, P. B. y Willis, S. L. (1982). En F. I. M. Craik & S. E. Trehub (Eds.), *Aging and cognitive processes* (pp. 353–389). New York: Plenum Press.

Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, NJ: Prentice-Hall.

Banks, S. (1990, May 3). Reprimands issued on test cheating. *Los Angeles Times,* p. B-3:1.

Barba, C. V. (1981). Mental development after dietary intervention: A study of Philippine children. *Journal of Cross-Cultural Psychology, 12,* 480–488.

Baumrind, D. (1993). The average expectable environment is not good enough: A response to Scarr. *Child Development, 64,* 1299–1317.

Bayley, N. y Oden, M. M. (1955). The maintenance of intellectual ability in gifted adults. *Journal of Gerontology, 10,* 91–107.

Beck, A. T. (1990). *Beck Anxiety Inventory manual.* San Antonio, TX: Psychological Corporation.

Beck, A. T. (1991). *Beck Scale for Suicide Ideation manual.* San Antonio, TX: Psychological Corporation.

Beck, A. T. y Steer, R. A. (1993). *Beck Depression Inventory: Manual.* San Antonio, TX: Psychological Corporation.

Bell, A. y Zubek, J. (1960). The effect of age on the intellectual performance of mental defectives. *Journal of Gerontology, 15,* 285–295.

Bellak, L. (1993). *The T.A.T., C.A.T., and S.A.T. in clinical use.* Des Moines, IA: Longwood Division, Allyn & Bacon.

Bellak, L. y Bellak, S. (1949). *Children's Apperception Test.* Larchmont, NY: C.P.S., Inc.

Bellak, L. y Bellak, S. (1973). *Manual: Senior Apperception Test.* Larchmont, NY: C.P.S., Inc.

Bellezza, F. S. y Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology, 16,* 151–155.

Bellezza, F. S. y Bellezza, S. F. (1995). Detection of copying on multiple-choice tests: An update. *Teaching of Psychology, 22,* 180–182.

Bem, D. J. y Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review, 81,* 506–520.

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting & Clinical Psychology, 42,* 165–172.

Bender, W. N. (1995). *Learning disabilities: Characteristics, identification, and teaching strategies* (2a. ed.). Boston: Allyn & Bacon.

Benjamin, L. T., Cavell, T. A. y Shallenberger, W. R. (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology, 11,* 133–141.

Ben-Porath, Y. S., Shondrick, D. D. y Stafford, K. P. (1995). MMPI-2 and race in a forensic diagnostic sample. *Criminal Justice and Behavior, 22,* 19–32.

Bergstrom, B. A. y Lunz, M. E. (1999). CAT for certification and licensure. En F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 67–91). Mahwah, NJ: Erlbaum.

Berliner, D. C. y Biddle, B. J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools.* Reading, MA: Addison-Wesley.

Berne, E. (1966). *Principles of group treatment.* New York: Oxford University Press.

Betsworth, D. G., Bouchard, T. J., Cooper, C. R., Grotevant, H. D., Hansen, J. C., Scarr, S. y Weinberg, R. A. (1994). Genetic and environmental influences on vocational interests assessed using adoptive and biological families and twins reared apart and together. *Journal of Vocational Behavior, 44,* 263–278.

Betz, N. E. (1992). Counseling uses of career self-efficacy theory. *Career Development Quarterly, 47*(1), 22–26.

Betz, N. E. (1994). Self-concept theory in career development and counseling. *Career Development Quarterly, 43,* 32–42.

Biemiller, L. (1986, enero 8). Critics plan assault on admissions tests and other standard exams. *Chronicle of Higher Education,* pp. 1, 4.

Binion, R. (1976). *Hitler among the Germans.* New York: Elsevier.

Black, H. (1962). *They shall not pass.* New York: Morrow.

Blakley, B. R., Quinones, M. A., Crawford, M. S. y Jago, I. A. (1994). The validity of isometric strength tests. *Personnel Psychology, 47,* 247–274.

Block, J. (1977). Recognizing the coherence of personality. En D. Magnusson & N. S. Endler (Eds.), *Interactional psychology: Current issues and future prospects.* New York: LEA/Wiley.

Bloom, B. S., Hastings, J. T. y Madaus, G. F. (1971). *Handbook of formative and summative evaluation of student learning.* New York: McGraw-Hill.

Bloom, B. S. y Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I. The cognitive domain.* New York: David McKay.

Blum, G. S. (1949). A study of the psychoanalytic theory of psychosexual development. *Genetic Psychology Monographs, 39,* 3–99.

Blum, G. S. (1950). *The Blacky Pictures.* New York: Psychological Corporation.

Bogardus, E. S. (1925). Measuring social distances. *Journal of Applied Sociology, 9,* 299–308.

Bond, L. (1989). The effects of special preparation on measures of scholastic ability. En R. L. Linn (Ed.), *Educational measurement* (3a. ed., pp. 429–444). New York: American Council on Education/Macmillan.

Borman, W. C., Hanson, M. y Hedge, J. (1997). Personnel selection. *Annual Review of Psychology, 48,* 299–337.

Bouchard, T. J., Jr., *et al*. (1983, junio). *Family resemblance for psychological interests.* Documento presentado en la reunión del International Congress on Twins Research, London.

Bouchard, T. J., Jr., Lykken, D. T., McGue, M., Segal, N. L. y Tellegen, A. (1990). Sources of human psychological differences: The Minnesota Study of Twins Reared Apart. *Science, 250,* 223–228.

Bouchard, T. J., Jr., & McGue, M. (1981). Familial studies of intelligence: A review. *Science, 212,* 1055–1059.

Bowman, M. L. (1989). Testing individual differences in Ancient China. *American Psychologist, 44,* 576–578.

Boyle, G. J. (1995). Review of the Personality Assessment Inventory. *Twelfth Mental Measurements Yearbook,* 764–766.

Boyle, M. H., Offord, D. R., Racine, Y. A., Szatmari, P., Sanford, M. y Fleming, J. E. (1996). Interviews versus checklists: Adequacy for classifying childhood psychiatric disorder based on adolescent reports. *International Journal of Methods in Psychiatric Research, 6,* 309–319.

Boyle, M. H., Offord, D. R., Racine, Y. A., Szatmari, P., Sanford, M. y Fleming, J. E. (1997). Adequacy of interviews vs. checklists for classifying childhood psychiatric disorder based on parent reports. *Archives of General Psychiatry, 54,* 793–799.

Braithwaite, V. A. y Scott, W. A. (1991). Values. En J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 661–753). San Diego, CA: Academic Press.

Brazelton, T. B. (1973). *Neonatal Behavioral Assessment Scale.* Philadelphia: Lippincott.

Brazelton, T. B. (1984). *Neonatal Behavioral Assessment Scale* (2nd ed.). Philadelphia: Lippincott.

Bredemeier, M. (1991). IQ test ban for blacks called unconstitutional. *California Association of School Psychologists Today,* nov./dic., 22–23.

Bricklin, B. (1984). *Bricklin Perceptual Scales.* Furlong, PA: Village.

Bridgman, C. S. y Hollenbeck, G. P. (1961). Effect of simulated applicant status on Kuder Form D occupational interest scores. *Journal of Applied Psychology, 45,* 237–239.

Brigham, C. C. (1923). *A study of American intelligence.* Princeton, NJ: Princeton University Press.

Brigham, C. C. (1930). Intelligence tests of immigrant groups. *Psychological Review, 37,* 158–165.

Brodie, F. M. (1983). *Richard Nixon: The shaping of his character.* Cambridge, MA: Harvard University Press.

Brody, N. (1992). *Intelligence* (2a. ed.). San Diego, CA: Academic Press.

Broman, S. H., Nichols, P. L., Shaughnessy, P. y Kennedy, W. (1987). *Retardation in young children.* Hillsdale, NJ: Erlbaum.

Bruvold, W. H. (1975). Judgmental bias in the rating of attitude statements. *Educational & Psychological Measurement, 45,* 605–611.

Bucholz, K. K., Marion, S. L., Shayka, J. J., Marcus, S. C. y Robins, L. N. (1996). A short computer interview for obtaining psychiatric diagnoses. *Psychiatric Services, 47,* 293–297.

Buck, J. N. (1992). *House–Tree–Person Projective Drawing Technique (H-T-P): Manual and interpretative guide* (revised by W. L. Warren). Los Angeles, CA: Western Psychological Services.

Bukatman, B. A., Foy, J. L. y De Grazia, E. (1971). What is competency to stand trial? *American Journal of Psychiatry, 127,* 1225–1229.

Bunderson, C. V., Inouye, D. K. y Olsen, J. B. (1989). The four generations of computerized educational measurement. En R. L. Linn (Ed.), *Educational measurement* (3a. ed., pp. 367–408). New York: Macmillan.

Bureau of Labor Statistics (1996). *Occupational outlook handbook.* Washington, DC: Author.

Bureau of Labor Statistics (2000). *Occupational outlook handbook, 2000–2001.* Washington, DC: Superintendent of Documents.

Burket, G. R. (1973). Empirical criteria for distinguishing and validating aptitude and achievement measures. En D. R. Green (Ed.), *The aptitude–achievement distinction.* Monterey, CA: CTB/McGraw-Hill.

Busse, E. W. y Maddox, G. (1985). *The Duke longitudinal studies of normal aging.* New York: Springer.

Butler, M., Retzlaff, P. y Vanderploeg, R. (1991). Neuropsychological test usage. *Professional Psychology: Research and Practice, 22,* 510–512.

Camara, W. J., Nathan, J. S. y Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research & Practice, 31,* 141–154.

Camara, W. J. y Schneider, D. L. (1994). Integrity tests: Facts and unresolved issues. *American Psychologist, 49,* 112–119.

Camara, W. J. y Schneider, D. L. (1995). Questions of construct breadth and openness of research in integrity testing. *American Psychologist, 50,* 459–460.

Camilli, G. y Shepard, L. A. (1994). *Methods for identifying biased test items.* Newbury Park, CA: Sage.

Campbell, D. P. (1965). A cross-sectional and longitudinal study of scholastic abilities over twenty-five years. *Journal of Counseling Psychology, 12,* 55–61.

Campbell, D. P. (1971). *Handbook for the Strong Vocational Interest Blank.* Stanford, CA: Stanford University Press.

Campbell, D. P. y Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Campbell, D. P. y Hansen, J. C. (1981). *Manual for the Strong–Campbell Interest Inventory* (3a. ed.). Stanford, CA: Stanford University Press.

Campbell, F. y Ramey, C. T. (1994). Effects of early intervention on intellectual and academic achievement: A follow-up study of children from low-income families. *Child Development, 65,* 684–698.

Campion, M. A., Pursell, E. D. y Brown, B. K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology, 41,* 25–42.

Canfield, A. A. (1951). The "sten" scale—a modified C scale. *Educational & Psychological Measurement, 11,* 295–297.

Cannell, J. J. (1988). Nationally normed elementary school testing in America's public schools: How all 50 states are testing above the national average (with commentaries). *Educational Measurement: Issues & Practice, 7*(2), 5–9.

Cannell, J. J. (1989). *How public educators cheat on achievement tests: The "Lake Wogebon" report.* Albuquerque, NM: Friends for Education.

Carlson, J. F. (1998). Review of the Beck Depression Inventory. *Thirteenth Mental Measurements Yearbook,* 117–120.

Carroll, J. B. (1973). The aptitude–achievement distinction: The case of foreign language aptitude and proficiency. En D. R. Green (Ed.), *The aptitude–achievement distinction.* Monterey, CA: CTB/McGraw-Hill.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Carson, A.D. (1998). Why has musical aptitude assessment fallen flat? And what can we do about it? *Journal of Career Assessment, 6,* 311–328.

Carver, R. P. (1974). Two dimensions of tests: Psychometric and edumetric. *American Psychologist, 29,* 512–518.

Cascio, W. F. (2000). Test utility. En A. E. Kazdin (Ed.), *Encyclopedia of psychology* (Vol. 8, pp. 52–55). Washington, DC: American Psychological Association.

Cascio, W. F., & Ramos, R. A. (1986). Development and application of new method for assessing job performance in behavioral economic terms. *Journal of Applied Psychology, 71,* 20–28.

Castro, J. G., & Jordan, J. E. (1977). Facet theory attitude research. *Educational Researcher, 6,* 7–11.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1–22.

Chaplin, W. F. y Goldberg, L. R. (1984). A failure to replicate the Ben and Allen study of individual differences in cross-situational consistency. *Journal of Personality and Social Psychology, 47,* 1074–1090.

Charles, D. C. y James, S. T. (1964). Stability of average intelligence. *Journal of Genetic Psychology, 105,* 105–111.

Chase, C. (1990–91). Essay test scoring: Expectancy and handwriting quality. *Psychology: A Journal of Human Behavior, 27*(4), 38–41.

Chauncey, H. y Dobbin, J. E. (1963). *Testing: Its place in education today.* New York: Harper & Row.

Chavez, S. (1993, August 19). SAT scores remain level in California. *Los Angeles Times,* pp. A1, A23.

Childs, A. y Klimoski, R. J. (1986). Successfully predicting career success: An application of the biographical inventory. *Journal of Applied Psychology, 71,* 3–8.

Chinn, P. C., Drew, C. J. y Logan, D. R. (1975). *Mental retardation: A life cycle approach.* St. Louis, MO: Mosby.

Christensen, H., Mackinnon, A., Jorm, A. F., Henderson, A. S., Scott, L. R. y Korten, S. E. (1994). Age differences and interindividual variation in cognition in community-dwelling elderly. *Psychology and Aging, 9,* 381–390.

Christenson, S. L. (1992). Review of the Child Behavior Checklist. *Eleventh Mental Measurements Yearbook,* 164–166.

Christiansen, K. y Knussman, R. (1987). Sex hormones and cognitive functioning in men. *Neuropsychobiology, 18,* 27–36.

Ciminero, A. R., Nelson, R. O. y Lipinski, D. P. (1977). Self-monitoring procedures. En A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment.* New York: Wiley.

Cocks, G. y Crosby, T. L. (Eds.). (1987). *Psycho/history: Readings in the method of psychology, psychoanalysis, and history.* New Haven, CT: Yale University Press.

Cohen, D. S., Colliver, J. A., Marcy, M. S., Fried, E. D. y Swartz, M. H. (1996). Psychometric properties of a standardized-patient checklist and rating-scale form used to assess interpersonal and communication skills. *Academic Medicine, 71* (Suppl. 1), S87–S89.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213–220.

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement, 10,* 237–255.

Cole, N. S. y Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3a. ed., pp. 201–219). New York: Macmillan.

College Entrance Examination Board. (1971). *Report of the Commission on Tests.* New York: Author.

Conners, C. K. (1973). Rating scales for use in drug studies with children. *Psychopharmacology Bulletin* [Special issue, Pharmacotherapy of children], 24–84.

Conners, C. K. y Barkley, R. A. (1985). Rating scales and checklists for child psychopharmacology. *Psychopharmacology Bulletin* [Special issue, Rating scales and assessment instruments for use in pediatric psychopharmacology research], *21,* 809–815.

Converse, P. E., Dotson, J. D., Hoag, W. J. y McGee III, W. H. (1980). *American social attitudes data sourcebook, 1947–78.* Cambridge, MA: Harvard University Press.

Cooley, H. H. (1922). *Human nature and the social order.* New York: Scribner's.

Cooper, J. B. y Pollock, D. (1959). The identification of prejudicial attitudes by the galvanic skin response. *Journal of Social Psychology, 50,* 241–245.

Corcoran, K. y Fischer, J. (2000). *Measures for clinical practice* (3a. ed., vols. 1 & 2). New York: Free Press.

Cordes, C. (1986, June). Test tilt: Boys outscore girls on both parts of the SAT. *APA Monitor,* pp. 30–31.

Costa, P. T., Jr. y McCrae, R. R. (1986). Personality stability and its implications for clinical psychology. *Clinical Psychology Review, 6,* 407–423.

Costantino, G. (1978, nov.). *Preliminary report on TEMAS: A new thematic apperception test to assess ego functions in ethnic minority children.* Documento presentado en la Second American Conference on Fantasy and the Imaging Process, Chicago.

Costantino, G., Malgady, R. y Rogler, L. H. (1988). *Tell-Me-A-Story—TEMAS—Manual.* Los Angeles: Western Psychological Services.

Courts, P. L. y McInerney, K. H. (1993). *Assessment in higher education: Politics, pedagogy, and portfolios.* Westport, CT: Praeger.

Crites, J. O. (1969). Interests. En R. L. Ebel (Ed.), *Encyclopedia of educational research* (4a. ed., pp. 678–685). New York: Macmillan.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3a. ed.). New York: Harper & Row.

Cronbach, L. J. y Drenth, P. J. D. (Eds.). (1972). *Mental tests and cultural adaptation.* The Hague: Mouton.

Cronbach, L. J. y Gleser, G. C. (1965). *Psychological tests and personnel decisions.* Urbana: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H. y Rajaratnam, N. (1972). *The dependability of behavioral measures: Theory of generalizability for scores and profiles.* New York: Wiley.

Cronin, J., Daniels, N., Hurley, A., Kroch, A. y Webber, R. (1975). Race, class, and intelligence: A critical look at the IQ controversy. *International Journal of Mental Health, 3*(4), 46–132.

Crowl, T. K. y McGinitie, W. H. (1974). The influence of students' speech characteristics on teachers' evaluations of oral answers. *Journal of Educational Psychology, 66,* 304–308.

Dahlstrom, W. G. y Gynther, M. D. (1986). Previous MMPI research on black Americans. En W. G. Dahlstrom, D. Lachar, & L. E. Dahlstrom (Eds.), *MMPI patterns of American minorities.* Minneapolis: University of Minnesota Press.

D'Amato, R. C. (1995). Review of the Adult Personality Inventory. *Twelfth Mental Measurements Yearbook,* 52–54.

Darley, J. B. y Hagenah, T. (1955). *Vocational interest measurement.* Minneapolis: University of Minnesota Press.

Das, J. P., Naglieri, J. A., & Kirby, J. P. (1994). *Assessment of cognitive processes: The PASS theory of intelligence.* Boston: Allyn & Bacon.

Davidshofer, C. (1985). Review of Jackson Vocational Interest Survey. *Ninth Mental Measurements Yearbook,* 739–740.

Debra v. *Turlington,* 644 F.2d 397 (1981); 730F.2d 1406 (1984).

Delis, D. C., & Jacobson, M. (2000). Neuropsychology: Testing. En A. E. Kazdin (Ed.), *Encyclopedia of psychology* (Vol. 5, pp. 423–430). New York: Oxford University Press.

Dember, W. N. (2001). The optimism–pessimism instrument: Personal and social correlates. En E. C. Chang (Ed.), *Optimism & pessimism: Implications for theory, research, and practice* (pp. 281–299). Washington, DC: American Psychological Association.

DeMille, R. (1962). Intellect after lobotomy in schizophrenia. *Psychological Monographs, 76*(16), 1–18.

Denton, L. (1988, August). Board votes to oppose Golden Rule technique. *APA Monitor,* p. 7.

Derogatis, L. R. (1994). *SCL-90-R: Symptom Checklist-90-R: Administration, scoring, and procedures manual* (3rd ed.). Minneapolis, MN: National Computer Systems.

Diamond, E. E. (1979). Sex equality and measurement practices. *New Directions for Testing and Measurement, 3,* 61–78.

Diana v. *State Board of Education,* C-70 37 RFT (N.D. Cal 1970).

Diekhoff, G. M. (1984). True–false tests that measure and promote structured understanding. *Teaching of Psychology, 11,* 99–101.

Dignon, A. M. (1996). Acceptability of a computer-administered psychiatric interview. *Computers in Human Behavior, 12,* 177–191.

Doebele, J. (1999, junio/julio). A common language: Community colleges become fluent in workforce development. *Community College Journal.*

Dolliver, R. H., Irvin, J. A. y Bigley, S. E. (1972). Twelve-year follow-up of the Strong Vocational Interest Blank. *Journal of Counseling Psychology, 19,* 212–217.

Donahue, D. y Sattler, J. M. (1971). Personality variables affecting WAIS scores. *Journal of Consulting & Clinical Psychology, 36,* 441.

Donlon, T. F. (Ed.). (1984). *The College Board technical handbook for the Scholastic Aptitude and achievement tests.* New York: College Entrance Examination Board.

Donnay, D. A. C. (1997). E. K. Strong's legacy and beyond: 70 years of the Strong Interest Inventory. *Career Development Quarterly, 46,* 2–22.

Doppelt, J. E. y Wallace, W. L. (1955). Standardization of the Wechsler Adult Intelligence Scale for older persons. *Journal of Abnormal and Social Psychology, 51,* 312–330.

Dorr-Bremme, D. W. y Herman, J. L. (1986). *Assessing student achievement: A profile of classroom practices* (CSE Monograph 11). Los Angeles: University of California, Center for the Study of Evaluation.

Dowd, E. T. (1992). Review of the Beck Hopelessness Scale. *Eleventh Mental Measurements Yearbook,* 81–82.

Dowd, E. T. (1998). Review of the Beck Anxiety Inventory. *Thirteenth Mental Measurements Yearbook,* 97–98.

Doyle, K. O., Jr. (1974). Theory and practice of ability testing in Ancient Greece. *Journal of the History of the Behavioral Sciences, 10,* 202–212.

Drake, R. M. (1954). *Drake Musical Aptitude Tests.* Chicago: Science Research Associates.

Drakeley, R. J., Herriot, P. y Jones, A. (1988). Biographical data, training success, and turnover. *Journal of Occupational Psychology, 61,* 145–152.

DuBois, P. H. (1970). *The history of psychological testing.* Boston: Allyn & Bacon.

Dudek, B. y Makowska, Z. (1993). Psychometric characteristics of the Orientation to Life Questionnaire for measuring the sense of coherence. *Polish Psychological Bulletin, 24,* 309–318.

Dunnette, M. D. (1963). Critics of psychological tests: Basic assumptions; how good? *Psychology in the Schools, 1,* 63–69.

Dunnette, M. D. y Borman, W. C. (1979). Personnel selection and classification systems. *Annual Review of Psychology, 30,* 477–525.

Dusky v. *United States,* 362 U.S. 402. (Abr. 18, 1960).

Dykens, E. M., Hodapp, R. M. y Leckman, J. F. (1994). *Behavior and development in fragile X syndrome.* Newbury Park, CA: Sage.

Ebel, R. L. (1979). *Essentials of educational measurement (3rd ed.).* Upper Saddle River, NJ: Prentice Hall.

Edelbrock, C. (1988). Informant reports. En E. S. Shapiro y T. R. Kratchowill (Eds.), *Behavioral assessment in schools: Conceptual foundations and practical applications* (pp. 351–383). New York: Guilford Press.

Edelbrock, C. y Achenbach, T. M. (1984). The teacher version of the Child Behavior Profile: 2. Boys aged 6–11. *Journal of Consulting & Clinical Psychology, 52,* 207–212.

Edens, J. F., Hart, S. D., Johnson, D. W., Johnson, J. K. y Olver, M. E. (2000). Use of the Personality Assessment Inventory to assess psychopathy in offender populations. *Psychological Assessment, 12,* 132–139.

Educational Testing Service. (1965). *ETS builds a test.* Princeton, NJ: Author.

Educational Testing Service. (1980a). *Test use and validity: A response to charges in the Nader/Nairn Report on ETS.* Princeton, NJ: Author.

Educational Testing Service (1980b). *Test scores and family income: A response to charges in the Nader-/Nairn Report on ETS.* Princeton, NJ: Author.

Educational Testing Service. (1992). *What we can learn from performance assessment for the professions.* ETS Conference on Education and Assessment. Princeton, NJ: Author.

Edwards, A. L. (1954). *Manual—Edwards Personal Preference Schedule.* New York: Psychological Corporation.

Egeland, B. (1985). Review of Wisconsin Card Sorting Test. En J. V. Mitchell (Ed.), *Ninth Mental Measurements Yearbook* (pp. 1746–1747). Lincoln: University of Nebraska Press.

Eisdorfer, C. (1963). The WAIS performance of the aged: A retest evaluation. *Journal of Gerontology, 18,* 169–172.

Ekman, P. y Friesen, W. V. (1978). *The Facial Action Coding System: A technique for the measurement of facial movement.* Palo Alto, CA: Consulting Psychologists Press.

Ekman, P. y Friesen, W. V. (1984). *Unmasking the face* (reprint ed.). Palo Alto, CA: Consulting Psychologists Press.

Ekstrom, R. B., French, J. W. y Harman, H. H. (1979). Cognitive factors: Their identification and replication. *Multivariate Behavior Research Monographs.* Ft. Worth, TX: Society for Multivariate Experimental Psychology.

Elam, S. M. (Ed.). (1978). *A decade of Gallup polls of attitudes toward education: 1969–1978.* Bloomington, IN: Phi Delta Kappa.

Elliott, S. N. y Busse, R. T. (1992). Review of the Child Behavior Checklist. *Eleventh Mental Measurements Yearbook,* 166–169.

Elms, A. (1976). *Personality and politics.* San Diego, CA: Harcourt Brace Jovanovich.

Erikson, E. H. (1969). *Gandhi's truth: On the origins of militant nonviolence.* New York: Norton.

Erikson, M. P. H. (1995). *Family centered assessment of young children at risk: The IDA readings.* Itasca, IL: Riverside Publishing.

Eron, L. (1950). A normative study of the TAT. *Psychological Monographs, 64* (Whole No. 315).

Esquivel, G. B. y Lopez, E. (1988). Correlations among measures of cognitive ability, creativity, and academic achievement for gifted minority children. *Perceptual and Motor Skills, 67,* 395–398.

Evans, W. (1984). Test wiseness: An examination of cue-using strategies. *Journal of Experimental Education, 52,* 141–144.

Exner, J. E. (1991). *The Rorschach: A comprehensive system. Vol 2. Current research and advanced interpretation.* New York: Wiley.

Exner, J. E. (1993). *The Rorschach: A comprehensive system. Vol. 1. Basic foundations* (3a ed.). New York: Wiley.

Eysenck, H. J. (1965). The effects of psychotherapy. *International Journal of Psychiatry, 1,* 97–178.

Eysenck, H. J. (1971). *The IQ argument.* New York: Library Press.

Eysenck, H. J. (Ed.). (1981). *A model for personality.* New York: Springer.

Eysenck, H. J. (1984). Recent advances in the theory and measurement of intelligence. *Early Child Development and Care, 15,* 97–115.

Fabiano, E. (1989). *Index to tests used in educational dissertations.* Phoenix, AZ: Oryx Press.

Farrell, A. D. (1993). Computers and behavioral assessment: Current applications, future possibilities, and obstacles to routine use. *Behavioral Assessment, 13,* 159–170.

Feather, N. T. (1986). Value systems across cultures: Australia and China. *International Journal of Psychology, 21,* 697–715.

Feigelson, M. E., y Dwight, S. A. (2000). Can asking questions by computer improve the candidness of responding? A meta-analytic perspective. *Consulting Psychology Journal: Practice & Research, 52,* 248–255.

Feldman, D. H. y Goldsmith, L. T. (1991). *Nature's gambit: Child prodigies and the development of human potential.* New York: Teachers College Press.

Fernandez, E. (1998). Review of the Beck Hopelessness Scale. *Thirteenth Mental Measurements Yearbook,* 123–125.

Feuerstein, R., Feuerstein, R. y Gross, S. (1997). The Learning Potential Assessment Device. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 297–313). New York: Guilford Press.

Fish, L. J. (1941). *One hundred years of examinations in Boston.* Dedham, MA: Transcript Press.

Fishbein, M. y Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research.* Reading, MA: Addison-Wesley.

Fisher, R. P., Geiselman, R. E., Raymond, D. S., Jurkevich, L. M. y Warhaftig, M. L. (1987). Enhancing enhanced eyewitness memory: Refining the cognitive interview. *Journal of Police Science and Administration, 15,* 201–297.

Fisher, R. P., McCauley, M. R. y Geiselman, R. E. (1994). Improving eyewitness testimony with the cognitive interview. En D. Ross, J. D. Read, & M. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 245–269). New York: Cambridge University Press.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327–358.

Flanagan, J. C., Tiedeman, D. V. y Willis, M. G. (1973). *The career data book.* Palo Alto, CA: American Institutes for Research.

Fleishman, E. A. (1972). On the relation between abilities, learning, and human performance. *American Psychologist, 27,* 1017–1032.

Fleishman, E. A. y Reilly, M. E. (1995). *Handbook of human abilities: Definitions, measurements, and job task requirements.* Potomac, MD: Management Research Institute.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171–191.

Flynn, J. R. (2000). The hidden history of IQ and special education: Can the problems be solved? *Psychology, Public Policy, & Law, 6,* 191–198.

Forbey, J. D., Handel, R. W. y Ben-Porath, Y. S. (2000). A real-data simulation of computerized adaptive administration of the MMPI-A. *Computers in Human Behavior, 16,* 83–96.

Forer, B. R. (1949). The fallacy of personal validation: A classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology, 44,* 118–123.

Fowler, R. D. (1966–1976). *Roche MMPI computerized interpretation service.* Nutley, NJ: Roche Psychiatric Institute.

Frank, L. K. (1939). Projective methods for the study of personality. *Journal of Psychology, 8,* 389–413.

Franklin, M. R., & Stillman, P. L. (1982). Examiner error in intelligence testing: Are you a source? *Psychology in the Schools, 19,* 563–569.

French, J. L. y Hale, R. L. (1990). A history of the development of psychological and educational testing. En C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence and achievement* (pp. 3–28). New York: Guilford Press.

Freud, S. (1905, reimpreso en 1959). Fragment of an analysis of a case of hysteria. En *Collected papers,* Vol 3. New York: Basic Books.

Freud, S. y Bullitt, W. C. (1967). *Thomas Woodrow Wilson.* Boston: Houghton Mifflin.

Frisby, C. L. (1999). Culture and test session behavior: Part II. *School Psychology Quarterly, 14,* 281–303.

Frueh, B. C., Smith, D. W. y Libet, J. M. (1996). Racial differences on psychological measures in combat veterans seeking treatment for PTSD. *Journal of Personality Assessment, 66,* 41–53.

Fulton, M., Thomson, G., Hunter, R., Raab, G., Laxen, D. y Hepburn, W. (1987). Influence of blood lead on the ability and attainment of children in Edinburgh. *Lancet, 1,* 1221–1226.

Funder, D. C. y Colvin, C. R. (1991). Some behaviors are more predictable than others. *The Score* (Newsletter of Division of the American Psychological Association), *13*(4), 3–4.

Gallup, G., Jr. (1991). *The Gallup Poll: Public opinions 1991.* Wilmington, DE: Scholarly Resources, Inc. (p. 92).

Galton, F. (1879). Psychometric experiments. *Brain, 2,* 149–162.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.

Gardner, H. (1997). Failing to act: Regrets of Terman's geniuses. *Journal of Creative Behavior, 31,* 120–124.

Gardner, W., Lidz, C. W., Mulvey, E. P. y Shaw, E. C. (1996). Clinical versus actuarial predictions of violence in patients with mental illnesses. *Journal of Counseling & Clinical Psychology, 64,* 602–609.

Geiger, M. A. (1990). Correlates of net gain from changing multiple-choice answers: Replication and extension. *Psychological Reports, 67,* 719–722.

Geiger, M. A. (1991a). Changing multiple-choice answers: Do students accurately perceive their performance? *Journal of Experimental Education, 59,* 250–257.

Geiger, M. A. (1991b). Changing multiple-choice answers: A validation and extension. *College Student Journal, 25,* 181–186.

Geiselman, R. E., Fisher, R. P., MacKinnon, D. P. y Holland, H. L. (1985). Eyewitness memory enhancement in the

police interview: Cognitive retrieval mnemonics versus hypnosis. *Journal of Applied Psychology, 70,* 401–412.

Georgia State Conferences of Branches of NAACP v. *State of Georgia.* Eleventh Circuit Court of Appeals, No. 84–8771 (1985).

Gerlach, V. S. y Sullivan, H. J. (1967). *Constructing statements of outcomes.* Inglewood, CA: Southwest Laboratory for Educational Research & Development.

Gerow, J. R. (1980). Performance on achievement tests as a function of the order of item difficulty. *Teaching of Psychology, 7,* 93–94.

Gesell, A. y Amatruda, C. S. (1941). *Developmental diagnosis.* New York: Paul B. Hoeber.

Getzels, J. W. y Jackson, P. W. (1962). *Creativity and intelligence: Explorations with gifted students.* New York: Wiley.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26,* 461–477.

Gifford, B. R. y O'Connor, M. C. (Eds.). (1992). *Changing assessments: Alternative views of aptitude, achievement, and instruction.* Boston: Kluwer.

Gill, K. y Keats, D. M. (1980). Elements of intellectual competence: Judgments by Australian and Malay university students. *Journal of Cross-Cultural Psychology, 11,* 233–243.

Glad, B. (1980). *Jimmy Carter: In search of the great White House.* New York: Norton.

Glass, G. V. y Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3a. ed.). Boston: Allyn & Bacon.

Glick, P., Gottesman, D. y Jolton, J. (1989). The fault is not in the stars: Susceptibility of skeptics and believers in astrology to the Barnum effect. *Personality and Social Psychology Bulletin, 15,* 572–583.

Glovrozov, P. A. (1974, July). Testing pupils orally. *Soviet Education, 16,* 95–105.

Glueck, B. C. y Reznikoff, M. (1965). Comparison of computer-derived personality profile and projective psychological test findings. *American Journal of Psychiatry, 121,* 1156–1161.

Goddard, H. H. (1920). *Human efficiency and levels of intelligence.* Princeton, NJ: Princeton University Press.

Goldberg, L. R. (1970). Man vs. model of man: A rationale, plus some evidence for a method of improving on clinical inferences. *Psychological Bulletin, 73,* 422–432.

Goldberg, L. R. (1980, April). *Some ruminations about the structure of individual differences: Developing a common lexicon for the major characteristics of human personality.* Documento presentado en la reunión anual de la Western Psychological Association, Honolulu, HI.

Goldman, B. A., Mitchell, D. F. y Egelson, P. E. (Eds.). (1997). *Directory of unpublished experimental mental measures* (Vol. 7). Washington, DC: American Psychological Association.

Goldstein, G. y Hersen, M. (1990). Historical perspective. En G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (2a. ed., pp. 3–17). New York: Pergamon.

Goodstadt, M. S. y Magid, S. (1977). When Thurstone and Likert agree: A confounding of methodologies. *Educational & Psychological Measurement, 37,* 811–818.

Gordon, E. (1965). The Musical Aptitude Profile: A new and unique musical aptitude test battery. *Council on Research in Musical Education,* No. 6, 12–16.

Gordon, R. y Peck, L. A. (1989). *The Custody Quotient.* Dallas, TX: Willington Institute.

Gottfredson, G. D., Holland, J. L. y Gottfredson, L. S. (1975). The relation of vocational aspirations and assessments to employment reality. *Journal of Vocational Behavior, 7,* 135–148.

Gottfredson, L. S. (1994). The science and politics of race-norming. *American Psychologist, 49,* 955–963.

Gottfredson, L. S. y Becker, H. J. (1981). A challenge to vocational psychology: How important are aspirations in determining male career development? *Journal of Vocational Behavior, 18,* 121–137.

Gough, H. G. y Bradley, P. (1996). *CPI manual* (3a. ed.). Palo Alto, CA: Consulting Psychologists Press.

Gould, S. J. (1981). *The mismeasure of man.* New York: Norton.

Granick, S. y Patterson, R. D. (1972). *Human aging, II: An eleven year follow-up biomedical and behavioral study.* Washington, DC: U.S. Government Printing Office.

Graves, M. (1948). *Design Judgment Test.* New York: Psychological Corporation.

Green, J. A. (1975). *Teacher-made tests* (2a. ed., pp. 122–135). New York: Harper & Row.

Green, K. (1984). Effects of item characteristics on multiple-choice item difficulty. *Educational & Psychological Measurement, 44,* 551–561.

Green, K. E. (1991). Measurement theory. En K. E. Green (Ed.), *Educational testing: Issues and applications* (pp. 3–25). New York: Garland Publishing.

Greene, H. A., Jorgensen, A. N. y Gerberich, J. R. (1954). *Measurement and evaluation in secondary school* (2a. ed.). New York: David McKay.

Greenfield, P. M. (1998). The cultural evolution of IQ. En U. Neisser (Ed.), *Intelligence on the rise?* Washington, DC: American Psychological Association.

Greenwald, A. G., McGhee, D. E. y Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality & Social Psychology, 74,* 1464–1480.

Griggs et al v. *Duke Power Company.* 401 U.S. 424, 3FEP175 (1971).

Gross, M. L. (1962). *The brain watchers.* New York: Random House.

Gross, M. L. (1965). Testimony before House Special Committee on Invasion of Privacy of the Committee on Government Operations. *American Psychologist, 20,* 958–960.

Grotevant, H. D., Scarr, S. y Weinberg, R. A. (1977). Patterns of interest similarity in adoptive and biological families. *Journal of Personality and Social Psychology, 35,* 667–676.

Guadalupe v. *Tempe Elementary School District,* Stipulation and Order (January 24, 1972).

Guilford, J. P. (1954). A factor analytic study across the domains of reasoning, creativity, and evaluation. I. Hypothesis and description of tests. *Reports from the Psychology Laboratory.* Los Angeles: University of Southern California.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Guilford, J. P. (1985). The structure-of-intellect model. En B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements and applications.* New York: Wiley.

Guilford, J. P. y Fruchter, B. (1973). *Fundamental statistics in psychology and education* (5a ed.). New York: McGraw-Hill.

Guttman, L. (1944). A basis for scaling quantitative data. *American Sociological Review, 9,* 139–150.

Gynther, M. D. (1981). Is the MMPI an appropriate assessment device for blacks? *Journal of Black Psychology, 7,* 67–75.

Haak, R. A. (1990). Using the sentence completion to assess emotional disturbance. En C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Personality, behavior, and context* (pp. 147–167). New York: Guilford Press.

Hack, M. y Breslau, N. (1985). Very low birth weight infants: Effects of brain growth during infancy on intelligence quotient at 3 years of age. *Pediatrics, 77,* 196–202.

Hager, P. (1991, Oct. 29). Court bans psychological tests in hiring. *Los Angeles Times,* p. A-20.

Haier, R. J. (1991). Cerebral glucose metabolism and intelligence. En P. A. Vernon (Ed.), *Biologic approaches to the study of human intelligence.* Norwood, NJ: Ablex.

Haladyna, T. M. y Downing, S. M. (1993). How many options is (sic) enough for a multiple-choice test item? *Educational & Psychological Measurement, 53,* 999–1010.

Hall, H. V. (1987). *Violence prediction: Guidelines for the forensic practitioner.* Springfield, IL: Charles C Thomas.

Hallahan, D. P., Kauffman, J. M. y Lloyd, J. W. (1996). *Introduction to learning disabilities.* Boston: Allyn & Bacon.

Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist, 52,* 1091–1101.

Hambleton, R. K. (1996). Advances in assessment models, methods, and practices. En D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 899–925). New York: Macmillan Reference.

Hambleton, R. K., Swaminathan, H. y Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hamersma, R. J., Paige, J. y Jordan, J. E. (1973). Construction of a Guttman facet designed cross-cultural attitude–behavior scale toward racial ethnic interaction. *Educational & Psychological Measurement, 33,* 565–576.

Hammer, E. G. y Kleiman, L. S. (1988). Getting to know you. *Personnel Administrator, 33*(5), 86–92.

Hammill, D. D., Brown, L., & Bryant, B. R. (1992). *A consumer's guide to tests in print* (2a. ed.). Austin, TX: pro.ed.

Hampson, E. (1990). Variations in sex-related cognitive abilities across the menstrual cycle. *Brain and Cognition, 14,* 26–43.

Hanes, K. R. (1998). Review of the Beck Scale for Suicide Ideation. *Thirteenth Mental Measurements Yearbook,* 125–126.

Haney, D. A. (1985, Feb. 3). Creative people: Their inner drive awes researchers. *Los Angeles Times,* I-2, 9.

Hanna, G. S. y Johnson, P. R. (1978). Reliability and validity of multiple-choice tests developed by four distractor selection procedures. *Journal of Educational Research, 71,* 203–206.

Hansen, J. C. (1984). The measurement of vocational interests: Issues and future directions. En R. B. Lent & S. D. Brown (Eds.), *Handbook of counseling psychology* (pp. 99–136). New York: Wiley.

Hansen, J. C. (1988). Changing interests of women: Myth or reality? *Applied Psychology: An International Review, 37*(2), 133–150.

Hansen, J. C. y Campbell, D. P. (1985). *Manual for the SVIB-SCII* (4a. ed.). Stanford, CA: Stanford University Press.

Harasty, J., Double, K. L., Halliday, G. M., Kril, J. J. y McRitchie, D. A. (1997). Language-associated cortical regions are proportionally larger in the female brain. *Archives of Neurology, 54,* 171–176.

Harmon, L. W., Hansen, J. C., Borgen, F. H. y Hammer, A. L. (1994). *Strong Interest Inventory: Applications and technical guide.* Palo Alto, CA: Consulting Psychologists Press.

Harrell, T. W. (1992). Some history of the Army General Classification Test. *Journal of Applied Psychology, 77,* 875–878.

Harrell, T. W. y Harrell, M. S. (1945). Army General Classification Test scores for civilian occupations. *Educational & Psychological Measurement, 5,* 229–342.

Harris, G. T. y Rice, M. E. (1996). The science in phallometric measurement of male sexual interest. *Current Directions in Psychological Science, 5,* 156–160.

Harris, M. M. y Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer–supervisor ratings. *Personnel Psychology, 41,* 43–62.

Harrow, A. J. (1972). *A taxonomy of the psychomotor domain: A guide for developing behavioral objectives.* New York: David McKay.

Hartshorne, H. y May, M. A. (1928). *Studies in the nature of character. Vol. 1: Studies in deceit.* New York: Macmillan.

Hathaway, S. R. y McKinley, J. C. (1989). *MMPI-2.* Minneapolis: University of Minnesota Press.

Hattie, J. (1980). Should creativity tests be administered under test-like conditions? An empirical study of three alternative conditions. *Journal of Educational Psychology, 72,* 87–98.

Hayes, D. P., Wolfer, L. T. y Wolfe, M. F. (1996). Schoolbook simplification and its relation to the decline in SAT-verbal scores. *American Educational Research Journal, 33,* 489–508.

Haynes, S. N. (1990). Behavioral assessment of adults. En G. Goldstein y M. Hersen (Eds.), *Handbook of psychological assessment* (2a. ed., pp. 423–463). New York: Pergamon.

Hays, J. R. (1997). Note on concurrent validity of the Personality Assessment Inventory in law enforcement. *Psychological Reports, 81,* 244–246.

Hebb, D. O. (1949). *The organization of behavior.* New York: Wiley.

Heilbrun, A. B., Jr. (1969). Parental identification and the patterning of vocational interests in college males and females. *Journal of Counseling Psychology, 16,* 342–347.

Herman, J. L. (1994). Item writing techniques. En T. Husén & T. N. Postlethwaite (Eds.), *International encyclopedia of education* (2a. ed., Vol. 5, pp. 3061–3066). Tarrytown, NY: Elsevier.

Herrnstein, R. J. y Murray, C. (1994). *The bell curve.* New York: The Free Press.

Hess, E. H. (1965). Attitude and pupil size. *Scientific American, 212,* 46–54.

Hess, E. H. (1975). *The tell-tale eye: How your eyes reveal hidden thoughts and emotions.* New York: Van Nostrand Reinhold.

Heubert, J. P. y Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion and graduation.* Washington, DC: National Research Council, National Academy Press.

Hier, D. B. y Crowley, W. F., Jr. (1982). Spatial ability in androgen-deficient men. *New England Journal of Medicine, 306,* 1202–1205.

Hirsch, N. D. M. (1926). A study of natio-racial mental differences. *Genetic Psychology Monographs, 1,* 231–406.

Hobbs, N. (1963). A psychologist in the Peace Corps. *American Psychologist, 18,* 47–55.

Hobson v. *Hansen,* 269 F. Suppl. 401 (D. D.C. 1967).

Hoffman, B. (1962). *The tyranny of testing.* New York: Crowell-Collier.

Hogan, J. y Quigley, A. (1994). Effects of preparing for physical ability tests. *Public Personnel Management, 23,* 85–104.

Holden, R. R., Fekken, G. C., Reddon, J. R., Helmes, E. y Jackson, D. N. (1988). Clinical reliabilities and validities of the Basic Personality Inventory. *Journal of Consulting & Clinical Psychology, 56,* 766–768.

Holland, J. L. (1985). *Making vocational choices: A theory of careers: A theory of vocational personalities and work environments* (2a. ed.). Upper Saddle River, NJ: Prentice Hall.

Holland, J. L. (1996). Exploring careers with a typology: What we have learned and some new directions. *American Psychologist, 51,* 397–406.

Holland, J. L. (1999). Why interest inventories are also personality inventories. En M. L. Savickas & A. R. Spokane (Eds.), *Vocational interests: Meaning, measurement, and counseling use* (pp. 87–101). Palo Alto, CA: Davies-Black Publishing/Consulting Psychologists Press.

Holmes, T. H. y Rahe, R. H. (1967). The Social Readjustment Scale. *Journal of Psychosomatic Research, 11,* 213–218.

Holt, A. (1974). *Handwriting in psychological interpretations.* Springfield, IL: Charles C. Thomas.

Holt, R. R. (1970). Yet another look at clinical and statistical prediction: Or, is clinical psychology worthwhile? *American Psychologist, 25,* 337–349.

Holtzman, W. H. (1988). Beyond the Rorschach. *Journal of Personality Assessment, 52,* 578–609.

Horn, C. A. y Smith, L. F. (1945). The Horn Art Aptitude Inventory. *Journal of Applied Psychology, 29,* 350–355.

Horn, J. L. (1982). The theory of fluid and crystallized intelligence in relation to concepts of cognitive psychology and aging in adulthood. En F. I. M. Craik & S. Trehub (Eds.), *Advances in the study of communication and affect: Volume 8: Aging and cognitive processes* (pp. 237–278). New York: Plenum.

Horn, J. L. y Hofer, S. M. (1992). Major abilities and development in the adult period. En R. J. Sternberg & C. A.

Berg (Eds.), *Intellectual development* (pp. 44–99). New York: Cambridge University Press.

Horn, J. M. (1983). The Texas Adoption Project: Adopted children and their intellectual resemblance to biological and adoptive parents. *Child Development, 54,* 268–275.

Howard, R. W. (2001). Searching the real world for signs of rising population intelligence. *Personality & Individual Differences, 30,* 1039–1058.

Hsu, T.-C., Moss, P. A. y Khampalikit, C. (1984). The merits of multiple-answer items as evaluated by using six scoring formulas. *Journal of Experimental Education, 52,* 152–158.

Hughes, H. H. y Converse, H. D. (1962). Characteristics of the gifted: A case for a sequel to Terman's study. *Exceptional Children, 29,* 178–183.

Hughes, S. (1995). Review of Denve: II. En J. C. Conoley y J. C. Impara (Eds.), *Twelfth Mental Measurements Yearbook* (pp. 263–265). Lincoln: Buros Institute of Mental Measures of the University of Nebraska– Lincoln.

Hunt, J. McV. (1961). *Intelligence and experience.* New York: Ronald Press.

Hunter, J. E. y Schmidt, F. L. (1990). *Methods of meta-analysis.* Newbury Park, CA: Sage.

Hunter, J. E. y Schmidt, F. L. (1996). Intelligence and job performance: Economic and social implications. *Psychology, Public Policy, & Law, 3,* 447–472.

Imada, A. S. (1982). Social interaction, observation, and stereotypes as determinants of differentiation in peer ratings. *Organizational Behavior & Human Performance, 29,* 397–415.

Impara, J. C. y Plake, B. S. (Eds.). (1998). *Thirteenth Mental Measurements Yearbook.* Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska, Lincoln.

Innocenti, G. M. (1994). Some new trends in the study of the corpus callosum. *Behavioral and Brain Research, 64,* 1–8.

International Assessment of Educational Progress. (1989). *A world of differences: An international assessment of math and science.* Princeton, NJ: Educational Testing Service.

Ireton, H. (1992). *Child Development Inventory: Manual.* Minneapolis, MN: Behavior Science Systems.

Ireton, H. (1998). *Preschool Development Inventory: Manual.* Minneapolis: Behavior Science Systems.

Isaacs, M., & Chen, K. (1990). Presence/absence of an observer in a word association test. *Journal of Personality Assessment, 55,* 41–51.

Jackson, D. N. (1998). *Multidimensional Aptitude Battery-II manual.* Port Huron, MI: Sigma Assessment Systems.

Jackson, D. N. (2000). *Jackson Vocational Interest Survey manual.* Port Huron, MI: Sigma Assessment Systems.

Jackson, D. N., Helmes, E., Hoffmann, H., Holden, R. R., Jaffe, P. G., Reddon, J. R. y Smiley, W. C. (1989). *Basic Personality Inventory manual.* Port Huron, MI: Sigma Assessment Systems.

Jackson, J. F. (1993). Human behavioral genetics, Scarr's theory, and her views on interventions: A critical review and commentary on their implications for African American children. *Child Development,* 64, 1318–1332.

Jackson, N. E. (1992). Precocious reading of English: Origins, structure, and predictive significance. In P. S. Klein & A. J. Tannenbaum (Eds.), *To be young and gifted* (pp. 171–203). Norwood, NJ: Ablex.

Jacobson, J. W. y Mullick, J. A. (1992). A new definition of mental retardation or a new definition of practice? *Psychology in Mental Retardation and Developmental Disabilities, 18,* 9–14.

Jamison, K. R. (1989). Mood disorders and patterns of creativity in British writers and artists. *Psychiatry, 52,* 125–134.

Jamison, K. R. (1993). *Touched with fire: Manic-depressive illness and the artistic temperament.* New York: Free Press.

Jancke, L. y Steinmetz, H. (1994). Interhemispheric-transfer time and corpus callosum size. *Neuroreport, 5,* 2385–2388.

Janos, P. M. y Robinson, N. M. (1985). Psychosocial development in intellectually gifted children. In F. D. Horowitz & M. O'Brien (Eds.), *The gifted and talented: Developmental perspectives*(pp. 149–195). Washington, DC: American Psychological Association.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39,* 1–123.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Jensen, A. R. (1981). *Straight talk about mental tests.* New York: Free Press.

Jensen, A. R., & Sinha, S. N. (1991). Physical correlates of human intelligence. In P. A. Vernon (Ed.), *Biological approaches to the study of human intelligence.* Norwood, NJ: Ablex.

Jessell, J. C. y Sullins, W. L. (1975). Effect of keyed response sequencing of multiple-choice items on performance and reliability. *Journal of Educational Measurement, 12,* 45–48.

Johnson, D. J. y Myklebust, H. R. (1967). *Learning disabilities: Educational principles and practices.* New York: Grune & Stratton.

Johnson, J. H. y Williams, T. (1975). The use of on-line computer technology in a mental health admitting system. *American Psychologist, 3,* 388–390.

Johnson, S. C., Pinkston, J. B., Bigler, E. D. y Blatter, D. D. (1996). Corpus callosum morphology in normal con-

trols and traumatic brain injury: Sex differences, mechanisms of injury, and neuropsychological correlates. *Neuropsychology, 10,* 408–415.

Joncas, J. y Standig, L. (1998). How much do accurate instructions raise scores on timed tests? *Perceptual & Motor Skills, 86,* 1257–1258.

Jones, H. E. y Conrad, H. S. (1933). The growth and decline of intelligence: A study of a homogeneous group. *Genetic Psychology Monographs, 13,* 223–298.

Jung, C. G. (1910). The association method. *American Journal of Psychology, 21,* 219–269.

Kaiser, S. y Wehrle, T. (1992). Automated coding of facial behavior in human computer interactions with FACS. *Journal of Nonverbal Behavior, 16,* 67–84.

Kansup, W. y Hakstian, A. R. (1975). Comparison of several methods of assessing partial knowledge in multiple-choice tests: Scoring procedures. *Journal of Educational Measurement, 12,* 219–230.

Kapes, J. T., Borman, C. A. y Frazier, N. (1989). An evaluation of the SIGI and DISCOVER microcomputer-based career guidance systems. *Measurement and Evaluation in Counseling and Development, 22,* 126–136.

Kapes, J. T. y Vansickle, T. R. (1992). Comparing paper–pencil and computer-based versions of the Harrington–O'Shea Career Decision Making System. *Measurement and Evaluation in Counseling and Development, 25,* 5–13.

Kaplan, H. I. y Sadock, B. J. (1995). *Comprehensive textbook of psychiatry* (6a. ed.). Baltimore: Williams & Wilkins.

Karp, S. A., Holmstrom, R. W. y Silber, D. E. (1990). *Apperceptive Personality Test Manual (Version 2.0).* Orland Park, IL: International Diagnostic Systems, Inc.

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., Williamson, D. y Ryan, N. (1997). Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime version (K-SADS-PL): Initial reliability and validity data. *Journal of the American Academy of Child & Adolescent Psychiatry, 36,* 980–988.

Kavan, M. G. (1995). Review of the Personality Assessment Inventory. *Twelfth Mental Measurements Yearbook,* 766–768.

Kazdin, A. E. (1998). *Research design in clinical psychology* (3a. ed.). Boston: Allyn & Bacon.

Kearns, D. (1976). *Lyndon Johnson and the American dream.* New York: Wilson.

Keating, D. P. (Ed.). (1976). *Intellectual talent: Research and development.* Baltimore, MD: Johns Hopkins University Press.

Kelly, E. L. y Fiske, D. W. (1951). *The prediction of performance in clinical psychology.* Ann Arbor: University of Michigan Press.

Kelly, G. A. (1955). *The psychology of personal constructs.* New York: Norton.

Kendall, P. C. y Norton-Ford, J. D. (1982). *Clinical psychology: Scientific and professional dimensions.* New York: Wiley.

Keyser, D. J. y Sweetland, R. C. (Eds.). (1984–1994). *Test critiques* (Vols. I–X). Austin, TX: pro.ed.

Kimura, D. y Hampson, E. (1993). Neural and hormonal mechanisms mediating sex differences in cognition. In P. A. Vernon (Ed.), *Biological approaches to the study of human intelligence* (pp. 375–397). Norwood, NJ: Ablex.

Kimura, D. y Hampson, E. (1994). Cognitive pattern in men and women is influenced by fluctuations in sex hormones. *Psychological Science, 3,* 57–61.

King, L. A. y King, D. W. (1993). *Sex-Role Egalitarianism Scale manual.* Port Huron, MI: Sigma Assessment Systems.

Kinicki, A. J. y Bannister, B. D. (1988). A test of the measurement assumptions underlying behaviorally anchored rating scales. *Educational & Psychological Measurement, 48,* 17–27.

Kirk, S. A., Gallagher, J. J. y Anastasiow, N. J. (1997). *Educating exceptional children* (8a. ed.). New York: Houghton Mifflin.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E. y Nizam, A. (1998). *Applied regression analysis and other multivariable methods* (3a. ed.). Pacific Grove, CA: Brooks/Cole.

Klimko, I. P. (1984). Item arrangement, cognitive entry characteristics, sex, and test anxiety as predictors of achievement examination performance. *Journal of Experimental Education, 52,* 214–219.

Klineberg, O. (1963). Negro–white differences in intelligence test performance. *American Psychologist, 18,* 198–203.

Knobloch, H. y Pasamanick, B. (Eds.). (1974). *Gesell and Amatruda's developmental diagnosis* (3rd ed.). New York: Harper & Row.

Knobloch, H., Stevens, F., & Malone, A. (1987). *Manual of developmental diagnosis: The administration and interpretation of the Revised Gesell and Amatruda Developmental and Neurological Examination.* Houston, TX: Developmental Evaluation Materials, Inc.

Kobak, A. A., Greist, J. H., Jefferson, J. W. y Katzelnick, D. J. (1996). Computer-administered clinical rating scales: A review. *Psychopharmacology, 127,* 291–301.

Kobak, K. A., Taylor, L. H., Dottl, S. L., Greist, J. H., Jefferson, J. W., Burroughs, D., Mantle, J. M., Katzelnick, D. J., Norton, R., Henk, H. J. y Serlin, R. C. (1997). A computer-administered telephone interview to identify mental disorders. *JAMA: Journal of the American Medical Association, 278,* 905–910.

Kohlberg, L. (1969). Stage and sequence: The cognitive–developmental approach to socialization. In D. Goslin (Ed.), *Handbook of socialization: Theory and research.* Chicago: Rand McNally.

Kohlberg, L. (1974). The development of moral stages: Uses and abuses. *Proceedings of the 1973 Invitational Conference on Testing Problems* (pp. 1–8). Princeton, NJ: Educational Testing Service.

Kohlberg, L. y Elfenbein, D. (1975). The development of moral judgments concerning capital punishment. *American Journal of Orthopsychiatry, 45,* 614–639.

Köhnken, G., Schimossek, E., Aschermann, E. y Höfer, E. (1995). The cognitive interview and the assessment of the credibility of adults' statements. *Journal of Applied Psychology, 80,* 671–684.

Korman, A. K. (1974). Disguised measure of civil rights attitudes. *Journal of Applied Psychology, 59,* 239–240.

Krathwohl, D. R., Bloom, B. S. y Masia, B. B. (1964). *Taxonomy of educational objectives: Handbook II, The affective domain.* New York: David McKay.

Kretschmer, E. (1925). *Physique and character.* New York: Harcourt Brace Jovanovich.

Krug, S. E. (1999). The Adult Personality Inventory. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (2a. ed., pp. 1211–1225). Mahwah, NJ: Erlbaum.

Kuder, G. F. (1963). A rationale for evaluating interests. *Educational & Psychological Measurement, 23,* 3–12.

Kuder, G. F. y Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2,* 151–160.

Kurtines, W. y Greif, E. B. (1994). The development of moral thought: Review and evaluation of Kohlberg's approach. En B. Puka (Ed.), *The great justice debate: Kohlberg criticism* (pp. 269–286). New York: Garland.

Lachar, D. (1999). Personality Inventory for Children, Second Edition (PIC-2), Personality Inventory for Youth (PIY), and Student Behavior Survey (SBS). En M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (2a. ed., pp. 399–427). Mahwah, NJ: Erlbaum.

Lacks, P. (1984). *Bender–Gestalt screening for brain dysfunction.* San Antonio, TX: The Psychological Corporation.

Lah, M. I. (1989). Sentence completion tests. En C. S. Newmark (Ed.), *Major psychological assessment instruments* (vol. 2, pp. 133–163). Boston: Allyn & Bacon.

Lancer, I. y Rim, Y. (1984). Intelligence, family size and sibling age spacing. *Personality & Individual Differences, 5,* 151–157.

Landauer, T. K. (1998). Learning and representing verbal meaning: The Latent Semantic Analysis Theory. *Current Directions in Psychological Science, 7,* 161–164.

Landauer, T. K. (1999). Latent semantic analysis: A theory of the psychology of language and mind. *Discourse Processes, 27,* 303–310.

Landers, S. (1989, Dec.). Test score controversy continues. *APA Monitor,* p. 10.

Langer, W. C. (1972). *The mind of Adolf Hitler.* New York: Basic Books.

Langevin, R. (1983). *Sexual strands: Understanding and treating sexual anomalies in men.* Hillsdale, NJ: Erlbaum.

Larry P. v. *Riles,* 495 F. Supp. 926 (N. D. Cal. 1979), appeal docketed, No. 80–4027 (9a. Cir., ene. 17, 1980).

Lee, E. S. (1951). Negro intelligence and selective migration: A Philadelphia test of the Klineberg hypothesis. *American Sociological Review, 16,* 227–233.

Lenke, J. M. (1988, Abril). Controversy fueled by district and state reports of achievement test results . . . "Lake Wobegon—or Not?" *The Score,* pp. 5, 13 (Newsletter of Division 5 of the American Psychological Association).

Lent, R. W., Lopez, F. G. y Bieschke, K. J. (1991). Mathematics self-efficacy: Sources and relation to science-based career choice. *Journal of Counseling Psychology, 4,* 424–430.

Leonard, C. M., Lombardino, L. J., Mercado, L. R., Browd, S. R., Breier, J. I. y Agee, O. F. (1996). Cerebral asymmetry and cognitive development in children: A magnetic resonance imaging study. *Psychological Science, 7,* 89–95.

Levine, M. (1976). The academic achievement test: Its historical context and social functions. *American Psychologist, 31,* 228–238.

Lewinsohn, P. M. (1965). Psychological correlates of overall quality of figure drawings. *Journal of Consulting Psychology, 29,* 504–512.

Lewis, M. y Jaskir, J. (1983). Infant intelligence and its relation to birth order and birth spacing. *Infant Behavior & Development, 6,* 117–120.

Liberman, R. P. (Ed.). (1988). *Psychiatric rehabilitation of chronic mental patients.* Washington, DC: American Psychiatric Press.

Liddell, D. L., Halpin, G. y Halpin, W. G. (1992). The Measure of Moral Orientation: Measuring the ethics of care and justice. *Journal of College Student Development, 33,* 325–330.

Lieberman, M. A. (1965). Psychological correlates of impending death: Some preliminary observations. *Journal of Gerontology, 20,* 71–84.

Lieberman, M. A. y Coplan, A. S. (1969). Distance from death as a variable in the study of aging. *Developmental Psychology, 2,* 71–84.

Lillienfeld, S. O., Alliger, G. y Mitchell, K. (1995). Why integrity testing remains controversial. *American Psychologist, 50,* 457–458.

Lindzey, G. (1965). Seer versus sign. *Journal of Experimental Research on Personality, 1,* 17–26.

Linn, R. L. (1992). Achievement testing. En M. C. Alkin (Ed.), *Encyclopedia of Educational Research* (6th ed., pp. 1–12). New York: Macmillan.

Lipsitt, P. D., Lelos, D. y McGarry, A. L. (1971). Competency for trial: A screening instrument. *American Journal of Psychiatry, 128,* 105–109.

Little, E. B. (1962). Overcorrection for guessing in multiple-choice test scoring. *Journal of Educational Research, 55,* 245–252.

Little, E. B. (1966). Overcorrection and undercorrection in multiple-choice test scoring. *Journal of Experimental Education, 35,* 44–47.

Lucas, A., Morley, R., Cole, T. J., Lister, G. y Leeson-Payne, C. (1992). Breast milk and subsequent intelligence quotient in children born preterm. *Lancet, 339,* 261–264.

Ludwig, A. M. (1995). *The price of greatness: Resolving the creativity and madness controversy.* New York: Guilford Press.

Lundeberg, M. A. y Fox, P. W. (1991). Do laboratory findings on test expectancy generalize to classroom outcomes? *Review of Educational Research, 61,* 94–106.

Lykken, D. T., Bouchard, T. J., McGue, M. y Tellegen, A. (1993). Heritability of interests: A twin study. *Journal of Applied Psychology, 78,* 649–661.

Lynn, R. (1982). IQ in Japan and the United States shows a growing disparity. *Science, 297,* 222–223.

Lynn, R. (1987). The intelligence of the mongoloids: A psychometric, evolutionary and neurological theory. *Personality and Individual Differences, 8,* 813–844.

Lynn, R. (1998). In support of the nutrition theory. En U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 207–215). Washington, DC: American Psychological Association.

Maccoby, E. E. y Maccoby, N. (1954). The interview: A tool of social science. En G. Lindzey (Ed.), *Handbook of social psychology* (pp. 449–487). Reading, MA: Addison-Wesley.

Machover, K. (1971). *Personality projection in the drawing of the human figure.* Springfield, IL: Charles C. Thomas.

MacKinnon, D. W. (1962). The nature and nurture of creativity talent. *American Psychologist, 17,* 484–495.

MacPhee, D., Ramey, C. T., & Yeates, K. O. (1984). Home environment and early cognitive development: Implications for intervention. En A. W. Gottfried (Ed.), *Home environment and early cognitive development. Longitudinal research.* Orlando, FL: Academic Press.

MacRae, H. M., Vu, N. V., Graham, B., Ward-Sims, M., Colliver, J. A. y Robbs, R. S. (1995). Comparing checklists and databases with physicians' ratings as measures of students' history and physical-examination skills. *Academic Medicine, 70,* 313–317.

Maddox, T. (Ed.). (1997). *Tests* (4a. ed.). Austin, TX: pro.ed.

Madhere, S. (1993). The development and validation of the Current Life Orientation Scale. *Psychological Reports, 72,* 467–472.

Maloney, D. P., Bouchard, T. J. y Segal, N. L. (1991). A genetic and environmental analysis of the vocational interests of monozygotic and dizygotic twins reared apart. *Journal of Vocational Behavior, 39,* 76–109.

Maloney, M. P. y Ward, M. P. (1976). *Psychological assessment: A conceptual approach.* New York: Oxford University Press.

Mantwill, M., Koehnken, G. y Aschermann, E. (1995). Effects of the cognitive interview on the recall of familiar and unfamiliar events. *Journal of Applied Psychology, 80,* 68–78.

Martin, E. y McDuffee, D. (1981). *A sourcebook of Harris national surveys: Repeated questions, 1963–76.* Chapel Hill: University of North Carolina, Institute for Research in Social Science.

Martorell, R. (1998). Nutrition and the worldwide rise in IQ scores. En U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 183–206). Washington, DC: American Psychological Association.

Maslow, A. H. (1954). *Motivation and personality.* New York: Harper & Row.

Masters, J. R. (1974). Relationship between number of response categories and reliability of Likert-type questionnaires. *Journal of Educational Measurement, 11,* 49–53.

Matarazzo, J. D. (1980). Behavioral health and behavioral medicine: Frontiers for a new health psychology. *American Psychologist, 35,* 807–817.

Matarazzo, J. D. (1992). Psychological testing and assessment in the 21st century. *American Psychologist, 47,* 1007–1018.

Maurer, S. D. y Fay, C. (1988). Effect of situational interviews, conventional structured interviews, and training on interview rating agreement: an experimental analysis. *Personnel Psychology, 41,* 329–344.

May, R. B. y Thompson, J. M. (1989). Test expectancy and question answering in prose processing. *Applied Cognitive Psychology, 3,* 261–269.

Mazlish, B. (1973). *In search of Nixon.* Baltimore, MD: Penguin.

McArthur, C. y Stevens, L. B. (1955). The validation of expressed interests as compared with inventoried interests: A fourteen-year follow-up. *Journal of Applied Psychology, 39,* 184–189.

McArthur, D. S. y Roberts, G. E. (1982). *Roberts Apperception Test for Children manual.* Los Angeles: Western Psychological Services.

McCall, R. B. (1979). The development of intellectual functioning in infancy and the prediction of later IQ. En J. D. Osofsky (Ed.), *Handbook of infant development* (pp. 707–741). New York: Wiley.

McCauley, M. R., & Fisher, R. P. (1995). Facilitating children's eyewitness recall with the revised cognitive interview. *Journal of Applied Psychology, 80,* 510–516.

McClelland, D. (1973). Testing for competence rather than for intelligence. *American Psychologist, 28,* 1–14.

McGarry, A. L., *et al*. (1973). *Competency to stand trial and mental illness.* Washington, DC: U.S. Government Printing Office.

McGue, M., Bouchard, T. J., Jr., Iacono, W. G. y Lykken, D. T. (1993). Behavioral genetics of cognitive ability: A life-span perspective. En R. Plomin & G. E. McClearn (Eds.), *Nature, nurture, and psychology* (pp. 59–76). Washington, DC: American Psychological Association.

McMichael, A. J., Baghurst, P. A., Wigg, N. R., Vimpani, G. V., Robertson, E. F. y Roberts, R. J. (1988). Port Pirie cohort study: Environmental exposure to lead and children's abilities at the age of four years. *New England Journal of Medicine, 319,* 468–475.

McNemar, Q. (1942). *The revision of the Stanford–Binet scale.* Boston: Houghton Mifflin.

McNemar, Q. (1964). Lost: Our intelligence? Why? *American Psychologist, 19,* 871–882.

McReynolds, P. (1986). History of assessment in clinical and educational settings. En R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 42–80). New York: Guilford Press.

Mead, A. D., & Drasgow, F. (1992). *Effects of administration: A meta-analysis.* Unpublished manuscript, University of Illinois, Champaign.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review, 69,* 1220–1232.

Meehl, P. E. (1954). *Clinical versus statistical prediction.* Minneapolis: University of Minnesota Press.

Meehl, P. E. (1965). Seer over sign: The first good example. *Journal of Experimental Research in Personality, 11,* 27–32.

Mehrabian, A. y Weiner, M. (1967). Decoding of inconsistent communication. *Journal of Personality and Social Psychology, 6,* 109–114.

Meier, N. C. (1942). *The Meier Art Tests. I. Art Judgment; Examiner's manual.* Iowa City: Bureau of Educational Research, University of Iowa.

Meijer, R. R. y Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23,* 223–237.

Meisels, S. J. y Fenichel, E. (Eds.). (1996). *New visions for the developmental assessment of infants and young children.* Itasca, IL: Riverside.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50,* 741–749.

Millman, J. y Pauk, W. (1969). *How to take tests.* New York: McGraw-Hill.

Millon, T., Millon, C., & Davis, R. (1994). *Manual for the MCMI-III.* Minneapolis, MN: NCS Assessments.

Mills, C. N. (1999). Development and introduction of a computer adaptive Graduate Record Examinations General Test. En F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 117–135). Mahwah, NJ: Erlbaum.

Mischel, W. (1968). *Personality and assessment.* New York: Wiley.

Mischel, W. (1986). *Introduction to personality* (4a. ed.). New York: Holt, Rinehart & Winston.

Mislevy, R. J. y Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57–75.

Moffatt, S. D. y Hampson, E. (1996). A curvilinear relationship between testosterone and spatial cognition in humans: Possible influence of hand preference. *Psychoneuroendocrinology, 21,* 323–337.

Molfese, V. J., DiLalla, L. F. y Bunce, D. (1997). Prediction of the intelligence test scores of 3- to 8-year-old children by home environment, socioeconomic status, and biomedical risks. *Merrill–Palmer Quarterly, 43,* 219–234.

Moreland, K. L., Eyde, L. D., Robertson, G. J., Primoff, E. S. y Most, R. B. (1995). Assessment of test user qualifications: A research-based measurement procedure. *American Psychologist, 50,* 14–23.

Morey, L. C. (1999). Personality Assessment Inventory. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (2nd ed., pp. 1083–1121). Mahwah, NJ: Erlbaum.

Mountain, M. y Snow, W. (1993). Wisconsin Card Sorting Test as a measure of frontal pathology: A review. *Clinical Neuropsychologist, 7,* 108–118.

Moyer, R. H. (1977). Environmental attitude assessment: Another approach. *Science Education, 61,* 347–356.

Murphy, K. R. y Davidshofer, C. O. (1994). *Psychological testing: Principles & applications* (3a. ed.). Upper Saddle River, NJ: Prentice Hall.

Murphy, L. L., Impara, J. C. y Plake, B. S. (Eds.). (1999). *Tests in print V.* Lincoln: The Buros Institute of Mental Measurements, the University of Nebraska, Lincoln.

Murray, B. (1998, August). The latest techno tool: Essay-grading computers. *APA Monitor, 29*(8), 43.

Murray, H. A. (and collaborators). (1938). *Explorations in personality.* New York: Oxford University Press.

Myart v. *Motorola,* 110 Cong. Record 5662–64 (1964).

Myers, I. B. y McCaulley, M. H. (1985). *Manual: A guide to the development and use of the Myers–Briggs Type Indicator.* Palo Alto, CA: Consulting Psychologists Press.

Nachmann, B. (1960). Childhood experiences and vocational choices in law, dentistry, and social work. *Journal of Counseling Psychology, 7,* 243–250.

Naglieri, J. A. y Das, J. P. (1997). *Das–Naglieri: Cognitive assessment system.* Itasca, IL: Riverside.

Naglieri, J. A., McNeish, T. y Bardos, A. (1991). *Draw-A-Person: Screening Procedure for Emotional Disturbance.* Austin, TX: pro.ed.

Naglieri, J. A. y Pfeiffer, S. I. (1992). Performance of disruptive behavior disordered and normal samples on the Draw A Person: Screening Procedure for Emotional Disturbance. *Psychological Assessment, 4,* 156–159.

Nairn, A. y Associates. (1980). *The reign of ETS: The corporation that makes up minds.* Washington, DC: Learning Research Project.

National Center for Education Statistics. (1996, noviembre). *Learning, curriculum, and achievement in international context.* Pittsburgh, PA: Superintendent of Documents.

National Center for Education Statistics. (1997, junio). *Pursuing excellence: A study of U.S. fourth-grade mathematics and science achievement in international context.* Pittsburgh, PA: Superintendent of Documents.

National Center for Education Statistics. (1998, feb.) *Pursuing excellence: A study of U.S. twelfth-grade mathematics and science achievement in international context.* Washington, DC: Author.

National Center for Education Statistics. (2001). *Digest of education statistics 2000.* Washington, DC: U.S. Department of Education.

National Center for Health Statistics. (1999). Births, marriages, divorces, and deaths for 1998. *Monthly Vital Statistics Report, 47*(21). Hyattsville, MD: The Center.

Needleman, H. L., Gunnoe, C., Leviton, A. y Perie, H. (1978). Neuropsychological dysfunction in children with "silent" lead exposure. *Pediatric Research, 12,* 1374. (Abstract).

Needleman, H. L., Schell, A., Bellinger, D., Leviton, A. y Allred, E. N. (1990). The long-term effects of exposure to low doses of lead in childhood. *New England Journal of Medicine, 322,* 83–88.

Nettler, G. (1959). Test burning in Texas. *American Psychologist, 14,* 682–683.

Nisbet, J. D. (1957). Intelligence and age: Retesting after twenty-four years' interval. *British Journal of Educational Psychology, 27,* 190–198.

Nixon, J. E. y Jewett, A. E. (1980). *An introduction to physical education* (9a. ed.). Philadelphia: Saunders.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3a. ed.). New York: McGraw-Hill.

Oakland, T. y Hu, S. (1993). International perspectives on tests with children and youths. *Journal of School Psychology, 31,* 501–517.

Ochse, R. (1991). The relation between creative genius and psychopathology: An historical perspective and a new explanation. *South African Journal of Psychology, 21,* 45–53.

Oden, M. H. (1968). The fulfillment of promise: 40-year follow-up of the Terman gifted group. *Genetic Psychology Monographs, 77,* 3–93.

Oliver, J. M., Cole, N. H., & Hollingsworth, H. (1991). Learning disabilities as functions of familial learning problems and developmental problems. *Exceptional Children, 57,* 427–440.

Olson, A. (2000). Computerized testing. *American School Board Journal, 187*(3), 31.

Ones, D. S. y Viswesvaran, C. (1998). Gender, age, and race differences on overt integrity tests: Results across four large-scale job applicant datasets. *Journal of Applied Psychology, 83*(1), 35–42.

Ones, D. S., Viswesvaran, C. y Schmidt, F. L. (1995). Integrity tests: Overlooked facts, resolved issues, and remaining questions. *American Psychologist, 50,* 456–457.

Ortar, G. (1963). Is a verbal test cross-cultural? *Scripta Hierosolymitana* (Hebrew University, Jerusalem), *13,* 329–335.

Osgood, C. E., Suci, G. J. y Tannenbaum, P. H. (1957). *The measurement of meaning.* Urbana: University of Illinois Press.

Osipow, S. H. (1983). *Theories of career development* (3a. ed.). New York: Appleton-Century-Crofts.

Ostrom, T. M., Bond, C. F., Jr., Krosnick, J. A. y Sedikides, C. (1994). Attitude scales: How we measure the unmeasureable. En S. Shavitt & T. C. Brock (Eds.), *Persuasion: Psychological insights and perspectives* (pp. 15–42). Boston: Allyn & Bacon.

Overall, J. E. y Gorham, D. R. (1962). The Brief Psychiatric Rating Scale. *Psychological Reports, 10,* 799–812.

Owen, S. V. (1992). Review of the Beck Hopelessness Scale. *Eleventh Mental Measurements Yearbook,* 82–83.

Owens, R. E., Hanna, G. S. y Coppedge, F. L. (1970). Comparison of multiple-choice tests using different types of distractor selection techniques. *Journal of Educational Measurement, 7,* 87–90.

Owens, W. A., Jr. (1953). Age and mental abilities: A longitudinal study. *Genetic Psychology Monographs, 48,* 3–54.

Owens, W. A., Jr. (1966). Age and mental abilities: A second adult follow-up. *Journal of Educational Psychology, 57,* 311–325.

Palmore, E. (1982). Predictors of the longevity difference: A 25-year follow-up. *Gerontologist, 225,* 513–518.

Palmore, E. y Cleveland, W. (1976). Aging, terminal decline, and terminal drop. *Journal of Gerontology, 31,* 76–86.

Parents in Action on Special Education (PASE) v. *Joseph P. Hannon,* No. 74C 3586 (N. D. III. 1980).

Paterson, D. G., Elliott, R. M., Anderson, L. D., Tooks, H. A. y Heidbreder, E. (1930). *The Minnesota Mechanical Ability Tests.* Minneapolis: University of Minnesota Press.

Paul, G. L. (1966). *Insight vs. desensitization in psychotherapy.* Stanford, CA: Stanford University Press.

Payne, A. F. (1928). *Sentence completions.* New York: New York Guidance Clinic.

Pedersen, N. L., Plomin, R., Nesselroade, J. R. y McClearn, G. E. (1992). A quantitative genetic analysis of cognitive abilities during the second half of the life span. *Psychological Science, 3,* 346–353.

Peterson, G. W., Ryan-Jones, R. E., Sampson, J. P., Readon, R. C., *et al*. (1994). A comparison of the effectiveness of three computer-assisted career guidance systems: Discover, SIGI, and SIGI PLUS. *Computers in Human Behavior, 10,* 189–198.

Peterson, R. C. y Thurstone, L. L. (1933). *Motion pictures and the social attitudes of children.* New York: Macmillan.

Piotrowski, C. (2000). How popular is the Personality Assessment Inventory in practice and training? *Psychological Reports, 86,* 65–66.

Piotrowski, C. y Keller, J. W. (1992). Psychological testing in applied settings: A literature review from 1982–1992. *Journal of Training & Practice in Professional Psychology, 6,* 74–82.

Pithers, W. D. y Laws, D. R. (1995). Phallometric assessment. In B. K. Schwartz & H. R. Cellini (Eds.), *The sex offender: Corrections, treatment and legal practice* (pp. 12–1 to 12–18). Kingston, NJ: Civic Research Institute.

Plake, B. S., Ansorge, C. J., Parker, C. S. y Lowry, S. R. (1982). Effects of item arrangement, knowledge of arrangement, test anxiety and sex on test performance. *Journal of Educational Measurement, 19,* 49–57.

Platt, J. R. (1961). On maximizing the information obtained from science examinations. *American Journal of Physics, 29,* 111–122.

Plomin, R. (1990). *Nature and nurture: An introduction to human behavior genetics.* Pacific Grove, CA: Brooks/Cole.

Plomin, R. y Foch, T. T. (1980). A twin study of objectively assessed personality in childhood. *Journal of Personality & Social Psychology, 39,* 680–688.

Posavec, E. J. y Carey, R. G. (1997). *Program evaluation* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Powers, D. E. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin, 100,* 67–77.

Powers, D. E. (1993). Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues & Practice, 12*(2), 24–30.

Powers, D. E. y Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning Test scores. *Journal of Educational Measurement, 36,* 93–118.

Prediger, D. J. y Hanson, G. R. (1976). Holland's theory of careers applied to men and women: Analysis of implicit assumptions. *Journal of Vocational Behavior, 8,* 167–184.

Preston, R. C. (1964). Ability of students to identify correct responses before reading. *Journal of Educational Research, 58,* 181–183.

Procter, M. (1993). Measuring attitudes. In N. Gilbert (Ed.), *Researching social life* (pp. 116–134). London: Sage.

Quay, H C. y Peterson, D. R. (1983). *Interim manual for the Behavior Problem Checklist.* Unpublished manuscript, University of Miami.

Raju, N. S., Normand, J. y Burke, M. J. (1990). A new approach for utility analysis. *Journal of Applied Psychology, 75,* 3–12.

Ramey, C. T., Campbell, F. A., Burchinal, M., Skinner, M. L., Gardner, D. M. y Ramely, S. L. (2000). Persistent effects of early childhood education on high-risk children and their mothers. *Applied Developmental Science, 4,* 2–14.

Randahl, G. J. (1991). A typological analysis of the relations between measured vocational interests and abilities. *Journal of Vocational Behavior, 38,* 333–350.

Rapaport, D., Gill, M. M. y Schafer, R. (1968). *Diagnostic psychological testing (rev. ed.).* New York: International Universities Press.

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from experiments. *Journal of Educational Psychology, 76,* 85–97.

Reilly, R. R. y Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35,* 1–62.

Reimanis, G. y Green, R. F. (1971). Imminence of death and intellectual decrement in the aging. *Developmental Psychology, 5,* 270–272.

Reise, S. P. y Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7,* 347–364.

Reitan, R. M. y Wolfson, D. (1993). The Halstead–Reitan *Neuropsychological Test Battery: Theory and clinical interpretation* (2a. ed.). Tucson, AZ: Neuropsychology Press.

Remmers, H. H. (1960). *Manual for the Purdue Master Attitude Scales.* Lafayette, IN: Purdue Research Foundation.

Reynolds, C. R., Chastain, R. L., Kaufman, A. S. y McLean, J. E. (1987). Demographic characteristics and IQ among adults: Analysts of the WAIS-R standardization sample as a function of the stratification variables. *Journal of School Psychology, 25,* 323–342.

Riegel, K. F. y Riegel, R. M. (1972). Development, drop, and death. *Developmental Psychology, 6,* 306–319.

Rieke, M. L. y Guastello, S. J. (1995). Unresolved issues in honesty and integrity testing. *American Psychologist, 50*(6), 458–459.

Roback, H. (1968). Human figure drawings: Their utility in the clinical psychologist's armamentarium for personality assessment. *Psychological Bulletin, 70,* 1–19.

Robbins, D. y Almond, E. (1992, Jan. 9). NCAA tightens academic rules for student-athletes. *Los Angeles Times,* pp. A1, A23.

Robinson, J. P., Shaver, P. R. y Wrightsman, L. S. (1991). *Measures of personality and social psychological attitudes.* New York: Academic Press.

Robinson, J. P., Shaver, P. R. y Wrightsman, L. S. (1999). *Measures of political attitudes. Measures of psychological attitudes* (vol. 2). San Diego, CA: Academic Press.

Robinson, N. M., Zigler, E. y Gallagher, J. J. (2000). Two tails of the normal curve: Similarities and differences in the study of mental retardation and giftedness. *American Psychologist, 55,* 1413–1424.

Rocklin, T. R., O'Donnell, A. M. y Holst, P. M. (1995). Effects and underlying mechanisms of self-adapted testing. *Journal of Educational Psychology, 87,* 103–116.

Rodgers, J. L., Cleveland, H. H, van den Oord, E. y Rowe, D. C. (2000). Resolving the debate over birth order, family size, and intelligence. *American Psychologist, 55,* 599–612.

Roe, A. (1956). *The psychology of occupations.* New York: Basic Books.

Roe, A. y Klos, D. (1969). Occupational classification. *Counseling Psychologist, 1,* 84–92.

Roe, A., & Siegelman, M. (1964). *The origin of interest.* Washington, DC: American Personnel and Guidance Association.

Rogers, C. R. y Dymond, R. F. (Eds.), (1954). *Psychotherapy and personality change*. Chicago: University of Chicago Press.

Rogers, R. y Shuman, D. W. (2000). *Conducting insanity evaluations* (2nd ed.). New York: Guilford Press.

Rogers, R., Ustad, K. L. y Salekin, R. T. (1998). Convergent validity of the Personality Assessment Inventory: A study of emergency referrals in correctional settings. *Assessment, 5,* 3–12.

Rogers, W. T. y Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational & Psychological Measurement, 59,* 234–247.

Rogers, W. T. y Yang, P.(1997). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment, 12,* 247–259.

Rokeach, M. (1968). *Beliefs, attitudes, and values: A theory of organization and change.* San Francisco: Jossey-Bass.

Rokeach, M. (1973). *The nature of human values.* New York: Free Press.

Rokeach, M. (1979). *Understanding human values.* Palo Alto, CA: Consulting Psychologists Press.

Rome, H. P., Swenson, W. M., Mataya, P., McCarthy, C. E., Pearson, J. S., Keating, F. R. y Hathaway, S. R. (1962). Symposium on automation techniques in personality assessment. *Proceedings of the Staff Meetings of the Mayo Clinic, 37,* 61–82.

Romer, D., Hornik, R., Stanton, B., Black, M., Li, X., Ricardo, I. y Feigelman, S. (1997). "Talking" computers: A reliable and private method to conduct interviews on sensitive topics with children. *Journal of Sex Research, 34,* 3–9.

Rose, L. C. y Gallup, A. M. (2001). The 33rd annual Phi Delta Kappa/Gallup Poll of the public's attendance toward the public schools. *Phi Delta Kappa, 83,* 41–48.

Rosenbaum, B. (1973). Attitude toward invasion of privacy in the personnel selection process and job applicant demographic and personality correlates. *Journal of Applied Psychology, 58,* 333–338.

Rosenman, R. H. (1986). Current and past history of Type A behavior pattern. En T. H. Schmidt, T. M. Dembroski, & G. Blumchen (Eds.), *Biological and psychological factors in cardiovascular disease* (pp. 15–40). New York: Springer-Verlag.

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L. y Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test.* Baltimore, MD: Johns Hopkins University Press.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom.* New York: Holt, Rinehart & Winston.

Rosenzweig, S. (1978). *Aggressive behavior and the Rosenzweig Picture-Frustration Study.* New York: Praeger.

Ross, C. C. y Stanley, J. C. (1954). *Measurement in today's schools* (3a. ed.). Upper Saddle River, NJ: Prentice Hall.

Rossi, P. H. y Freeman, H. E. (1993). *Evaluation: A systematic approach* (5a. ed.). Beverly Hills, CA: Sage.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A. y Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology, 75,* 175–184.

Rotter, J. B. (1954). *Social learning and clinical psychology.* Upper Saddle River, NJ: Prentice Hall.

Rotter, J. B., Lah, M. I. y Rafferty, J. E. (1992). *Rotter Incomplete Sentences Blank manual.* San Antonio, TX: Psychological Corporation.

Rourke, B. P. (Ed.). (1989). *Nonverbal learning disabilities: The syndrome and the model.* New York: Guilford Press.

Rowley, G. L. (1974). Which examinees are most favoured by the use of multiple-choice tests? *Journal of Educational Measurement, 11,* 15–23.

Rudman, L. A., Greenwald, A. G., Mellott, D. S. y Schwartz, J. L. K. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the Implicit Association Test. *Social Cognition, 17,* 437–465.

Russell, M. y Karol, D. (1994). *The 16 PF Fifth Edition administrator's manual.* Champaign, IL: Institute for Personality and Ability Testing.

Ryan, J., Prefitera, A. y Powers, L. (1983). Scoring reliability on the WAIS-R. *Journal of Consulting & Clinical Psychology, 51,* 149–150.

Sattler, J. M. (1988). *Assessment of children* (3a. ed.). San Diego, CA: Jerome M. Sattler.

Sattler, J. M., Hillix, W. A., & Neher, L. A. (1970). Halo effect in examiner scoring of intelligence test responses. *Journal of Consulting & Clinical Psychology, 34,* 172–176.

Sattler, J. M. y Winget, B. M. (1970). Intelligence testing procedures as affected by expectancy and IQ. *Journal of Clinical Psychology, 26,* 446–448.

Savitz, F. R. (1985). Effects of easy examination questions placed at the beginning of science multi-choice examinations. *Journal of Instructional Psychology, 12,* 6–10.

Scarr, S. (1992). Developmental theories for the 1990s: Development and individual differences. *Child Development, 63,* 1–19.

Scarr, S. (1993). Biological and cultural diversity: The legacy of Darwin for development. *Child Development, 64,* 1333–1353.

Scarr, S. y Weinberg, R. A. (1983). How people make their own environments: A theory of genotype–environment effects. *Child Development, 54,* 424–435.

Schaie, K. W. (1990). The optimization of cognitive functioning in old age: Prediction based on cohort-sequential and longitudinal data. En P. B. Baltes & M. Baltes (Eds.), *Longitudinal research and the study of successful (optimal) aging* (pp. 94–117). New York: Cambridge University Press.

Schaie, K. W. (1994). The course of adult intellectual development. *American Psychologist, 49,* 304–313.

Schaie, K. W. y Hertzog, C. (1983). Fourteen-year cohort-sequential analyses of adult intellectual development. *Developmental Psychology, 19,* 531–543.

Schaie, K. W. y Willis, S. L. (1986). Can decline in adult cognitive functioning be reversed? *Developmental Psychology, 22,* 223–232.

Scheuneman, J. D. y Bleistein, C. A. (1989). A consumers' guide to statistics for identifying differential item functioning. *Applied Measurement in Education, 2,* 255–275.

Schinke, S. (1995). Review of the Eating Disorder Inventory-2. *Twelfth Mental Measurements Yearbook,* 333–335.

Schlaug, G., Jaencke, L., Huang, Y, Staiger, J. F., *et al.* (1995). Increased corpus callosum size in musicians. *Neuropsychologia, 33,* 1047–1055.

Schlaug, G., Jaencke, L., Huang, Y. y Steinmetz, H. (1995). In vivo evidence of structural brain asymmetry in musicians. *Science, 267* (5198), 699–701.

Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, H. R., Pearlman, K. y McDaniel, M. (1993). Refinements in validity generalization methods: Implications for the situational specificity hypothesis. *Journal of Applied Psychology, 78,* 3–12.

Schmidt, F. L., Ones, D. S. y Hunter, J. E. (1992). Personnel selection. *Annual Review of Psychology, 43,* 627–670.

Schmidt, S. R. (1983). The effects of recall and recognition test expectancies on the retention of prose. *Memory and Cognition, 11,* 172–180.

Schmitt, N. y Robertson, I. (1990). Personnel selection. *Annual Review of Psychology, 41,* 289–391.

Schneider, M. F. (1989). *Children's Apperceptive Storytelling Test.* Austin, TX: pro.ed.

Schneider, M. F. y Perney, J. (1990). Development of the Children's Apperceptive Story-telling Test. *Psychological Assessment: A Journal of Consulting & Clinical Psychology, 2,* 179–185.

Schoenfeldt, L. F. y Mendoza, J. L. (1994). Developing and using factorially derived biographical scales. En G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 147–169). Palo Alto, CA: Consulting Psychologists Press.

Schwab, D. P. y Packard, G. L. (1973). Response distortion on the Gordon Personal Inventory and the Gordon Personal Profile in the selection context: Some implications for predicting employee behavior. *Journal of Applied Psychology, 58,* 372–374.

Schweinhart, L. y Weikart, D. (1997). The High Scope preschool curricular comparison study through age 23. *Early Childhood Research Quarterly, 12,* 117–143.

Scribner, S. y Cole, M. (1973). Cognitive consequences of formal and informal schooling. *Science, 182,* 553–559.

Scully, J. A., Tosi, H. y Banning, K. (2000). Life event checklists: Revisiting the Social Readjustment Ra-

ting Scale after 30 years. *Educational & Psychological Measurement, 60,* 864–876.

Sears, R. R. (1977). Sources of life satisfactions of the Terman gifted men. *American Psychologist, 32,* 119–128.

Seashore, C. E. (1939). *Psychology of music.* New York: McGraw-Hill.

Segall, D. O. y Moreno, K. E. (1999). Development of the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery. En F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 35–65). Mahwah, NJ: Erlbaum.

Sellin, T. y Wolfgang, M. E. (1964). *The measurement of delinquency.* New York: Wiley.

Selltiz, C., Wrightsman, L. S. y Cook, S. W. (1976). *Research methods in social relations* (3a. ed.). New York: Holt, Rinehart & Winston.

Serlin, R. C. y Kaiser, H. F. (1978). Method for increasing the reliability of a short multiple-choice test. *Educational & Psychological Measurement, 38,* 337–340.

Shaha, S. H. (1984). Matching test: Reduced anxiety and increased test effectiveness. *Educational & Psychological Measurement, 44,* 869–881.

Shaywitz, B. A., Shaywitz, S. E., Pugh, K. R., Constable, R. T., Skudlarski, P., Fulbright, R. K., Bronen, R. A., Fletcher, J. M., Shankweller, D. P., Katz, L. y Gore, J. C. (1995). Sex differences in the functional organization of the brain for language. *Nature, 373,* 607–609.

Shea, C. (1994, September 7). "Gender gap" on examinations shrank again this year. *Chronicle of Higher Education,* p. A54.

Sheldon, W. H., Stevens, S. S., & Tucker, W. B. (1940). *The varieties of human physique.* New York: Harper & Row.

Shogren, E. (1997, septiembre 16). Debate over national school tests offers real-life lesson in politics. *Los Angeles Times,* p. A5.

Siegelman, M. y Peck, R. F. (1960). Personality patterns related to occupational roles. *Genetic Psychology Monographs, 61,* 291–349.

Siegler, I. C., McCarty, S. M. y Logue, P. E. (1982). Wechsler Memory Scale scores, selective attribution, and distance from death. *Journal of Gerontology, 37,* 176–181.

Sigman, M. y Whaley, S. E. (1998). The role of nutrition in the development of intelligence. En U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 155–182). Washington, DC: American Psychological Association.

Silva, F., Martinez, A., Moro, M. y Ortet, G. (1996). Dimensions of interpersonal orientation: Description and construct validation of the Spanish assessment kit. *European Psychologist, 1,* 187–199.

Silverman, L. K. (1995). Highly gifted children. En J. Genschaft, M. Birely, & C. Hollinger (Eds.), *Serving gifted and talented students* (pp. 217–240). Austin, TX: pro.ed.

Simpson, E. J. (1966). The classification of educational objectives, psychomotor domain. *Illinois Teacher of Home Economics, 10,* 110–114.

Sinacore, J. M., Connell, K. J., Olthoff, A. J., Friedman, M. H. y Gecht, M. R. (1999). A method for measuring interrater agreement on checklists. *Evaluation & the Health Professions, 22,* 221–234.

Sines, J. O. (1970). Actuarial versus clinical prediction in psychopathology. *British Journal of Psychiatry, 116,* 129–144.

Slack, W. V. y Porter, D. (1980). The Scholastic Aptitude Test: A critical appraisal. *Harvard Educational Review, 50,* 154–175.

Slate, J. R. y Jones, C. H. (1990). Student error in administering the WISC-R: Identifying problem areas. *Measurement and Evaluation in Counseling and Development, 23,* 137–140.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47,* 149–155.

Snyder, C. R. (1974). Acceptance of personality interpretations as a function of assessment procedures. *Journal of Consulting Psychology, 42,* 150.

Snyderman, M. y Rothman, S. (1987). Survey of expert opinion on intelligence and aptitude testing. *American Psychologist, 42,* 137–144.

Sobel, D., & Andrewes, W. J. H. (1998). *The illustrated longitude.* New York: Walker & Co.

Society for Industrial and Organizational Psychology, Inc. (1987). *Principles for the validation and use of personal selection procedures* (3a. ed.). College Park, MD: Author.

Sokal, M. M. (Ed.). (1987). *Psychological testing and American society 1890–1930.* New Brunswick, NJ: Rutgers University Press.

Soroka v. *Dayton-Hudson Corp.* 91. L.A. Daily Journal D.A.R. 13204 (Cal. Ct. App. 1991).

Spearman, C. E. (1927). *The abilities of man.* London: Macmillan.

Speath, J. L. (1976). Characteristics of the work setting and the job as determinants of income. En W. H Sewell, R. M. Sauser, & D. L. Featherman (Eds.), *Schooling and achievement in American society.* New York: Academic Press.

Spohr, H. y Steinhausen, H. (Eds.). (1996). *Alcohol, pregnancy, and the developing child.* New York: Cambridge University Press.

Stamoulis, D. T. y Hauenstein, N. M. A. (1993). Rater training and rating accuracy: Training for dimensional

accuracy versus training for ratee differentiation. *Journal of Applied Psychology, 78,* 994–1003.

Stanford, G. y Oakland, T. (2000). Cognitive deficits underlying learning disabilities: Research perspectives from the United States. *School Psychology International, 21,* 306–321.

Stanley, J. C., Keating, D. P. y Fox, L. H. (Eds.). (1974). *Mathematical talent: Discovery, description, and development.* Baltimore, MD: Johns Hopkins University Press.

Starch, D. y Elliott, E. C. (1913). Reliability of grading work in mathematics. *School Review, 21,* 254–259.

Steelman, L. C. y Doby, J. T. (1983). Family size and birth order as factors on the IQ performance of black and white children. *Sociology of Education, 56,* 101–109.

Steimel, R. J. y Suziedelis, A. (1963). Perceived parental influence and inventoried interests. *Journal of Counseling Psychology, 10,* 289–295.

Stell v. *Savannah–Chatham County Board of Education.* 210 F Supp. 667, 668 (S.D. Ga. 1963), rev'd 333 F.2d 55 (5a. Cir. 1964), cert. den. 379 U.S. 933 (1964).

Stephenson, W. (1953). *The study of behavior: Q-technique and its methodology.* Chicago: University of Chicago Press.

Sternberg, C. (1955). Personality trait patterns of college students majoring in different fields. *Psychological Monographs, 69,* No. 18 (Whole No. 403).

Sternberg, R. J. (1982). Thinking and learning skills: A view of intelligence. *Education Digest, 47,* 20–22.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence.* New York: Cambridge University Press.

Sternberg, R. J. (1986). *The triarchic mind: A new theory of human intelligence.* New York: Viking.

Sternberg, R. J. (1988). Mental self-government: A theory of intellectual styles and their development. *Human Development, 31,* 197–224.

Stewart, J. R. (1998). Review of the Beck Scale for Suicide Ideation. *Thirteenth Mental Measurements Yearbook,* 126–127.

Stewart, L. H. (1959). Mother–son identification and vocational interest. *Genetic Psychology Monographs, 60,* 31–63.

Stokes, G. S., Mumford, M. D., & Owens, W. A. (1994). *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction.* Palo Alto, CA: Consulting Psychologist Press Books.

Stott, D. H. (1983). Brain size and "intelligence." *British Journal of Developmental Psychology, 1,* 279–287.

Strang, H. R. (1980). Effect of technically worded options on multiple-choice test performance. *Journal of Educational Research, 73,* 262–265.

Streissguth, A., Bookstein, F. y Barr, H. (1996). A dose–response study of the enduring effects of prenatal alcohol exposure: birth to 14 years. En H. Spohr & H. Steinhausen (Eds.), *Alcohol, pregnancy, and the developing child.* New York: Cambridge University Press.

Strong, E. K., Jr. (1955). *Vocational interests 18 years after college.* Minneapolis: University of Minnesota Press.

Sullivan, G. S., Mastropieri, M. A. y Scruggs, T. E. (1995). Reasoning and remembering: Coaching students with learning disabilities to think. *Journal of Special Education, 29,* 310–322.

Sulsky, L. M. y Day, D. V. (1994). Effects of frame-of-reference training on rater accuracy under alternative time and delays. *Journal of Applied Psychology, 79,* 515–543.

Sundberg, N. D. (1977). *Assessment of persons.* Upper Saddle River, NJ: Prentice Hall.

Super, D. E. (1973). The Work Values Inventory. En D. G. Zytowski (Ed.), *Contemporary approaches to interest measurement.* Minneapolis: University of Minnesota Press.

Super, D. E. y Bohn, M. J., Jr. (1970). *Occupational psychology.* Belmont, CA: Wadsworth.

Super, D. E. y Crites, J. O. (1962). *Appraising vocational fitness.* New York: Harper & Row.

Supple, A. J., Aquilino, W. S. y Wright, D. L. (1999). Collecting sensitive self-report data with laptop computers: Impact on the response tendencies of adolescents in a home interview. *Journal of Research on Adolescence, 9,* 467–488.

Swanson, J. L. (1993). Integrated assessment of vocational interests and self-rated skills and abilities. *Journal of Career Assessment,* 1, 50–65.

Swartz, J. D. (1992). The HIT and HIT 25: Comments and clarifications. *Journal of Personality Assessment, 58,* 432–433.

Swenson, W. M. y Pearson, J. S. (1964). Automation techniques in personality assessment: A frontier in behavioral science and medicine. *Methods of Information in Medicine, 3,* 34–36.

Swinton, S. S. y Powers, D. E. (1985). *The impact of self-study on GRE test performance* (Res. Rep. 85–12). Princeton, NJ: Educational Testing Service.

Tarasoff v. *Regents of University of California,* 17 Cal. 3d 425 (1983).

Taylor, H. C. y Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565–578.

Taylor, J. A. (1953). A personality scale of manifest anxiety. *Journal of Abnormal and Social Psychology, 48,* 285–290.

Taylor, R. G. y Lee, E. (1995). A review of the methods and problems of measuring reliability for criterion referenced tests and items. *Journal of Instructional Psychology, 22,* 88–94.

Teeter, P. A. (1985). Review of Adjective Check List. *Ninth Mental Measurements Yearbook,* 50–52.

Tenopyr, M. L. (1996). The complex interaction between measurement and national employment policy. *Psychology, Public Policy, and Law, 2,* 348–362.

Terman, L. M. y Merrill, M. A. (1973). *Stanford–Binet Intelligence Scale: 1972 norms edition.* Boston: Houghton Mifflin.

Terman, L. M. y Oden, M. H. (1959). *The gifted group at mid-life. Genetic studies of genius. V.* Stanford, CA: Stanford University Press.

Thatcher, R. W., Lester, M. L., McAlaster, R., Horst, R. y Ignasias, S. W. (1983). Intelligence and lead toxins in rural children. *Journal of Learning Disabilities, 16,* 355–359.

Thomas, G. E., Alexander, K. L. y Eckland, B. K. (1979). Access to higher education: The importance of race, sex, social class, and academic credentials. *School Review, 87,* 133–156.

Thomas, R. G. (1985). Review of Jackson Vocational Interest Survey. *Ninth Mental Measurements Yearbook,* 740–742.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54,* 837–847.

Thompson, C. (1949). The Thompson modification of the Thematic Apperception Test. *Journal of Projective Techniques, 13,* 469–478.

Thoreson, C. E. y Mahoney, M. J. (1974). *Behavioral self-control.* New York: Holt, Rinehart & Winston.

Thorndike, E. L. (1912). The permanence of interests and their relation to abilities. *Popular Science Monthly, 81,* 449–456.

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement, 8,* 63–70.

Thorndike, R. L., Hagen, E. P. y Sattler, J. M. (1986). *The Stanford–Binet Intelligence Scale: Fourth Edition, Technical manual.* Chicago: Riverside.

Tidwell, R. (1980). Biasing potential of multiple-choice test distractors. *Journal of Negro Education, 49,* 280–296.

Tittle, C. K. (1984). Test bias. En T. N. Husén & T. Postlethwaite (Eds.), *International encyclopedia of education* (pp. 5199–5204). New York: Wiley.

Tokar, D. M. y Fischer, A. R. (1998). More of RIASEC and five-factor model of personality: Direct assessment of Prediger's (1982) and Hogan's (1983) dimensions. *Journal of Vocational Behavior, 56,* 246–255.

Tokar, D. M. y Swanson, J. L. (1995). Evaluation of the correspondence between Holland's vocational personality typology and the five factor model of personality. *Journal of Vocational Behavior, 46,* 89–108.

Tombari, M. y Borich, G. (1999). *Authentic assessment in the classroom: Applications and practice.* Upper Saddle River, NJ: Prentice Hall.

Torrance, E. P. (1988). The nature of creativity as manifest in its testing. En R. J. Sternberg (Ed.), *The nature of creativity: Contemporary psychological perspectives.* New York: Cambridge University Press.

Trull, T. J., Widiger, T. A., Useda, J. D., Holcomb, J., Doan, B. T., Axelrod, S. R., Stern, B. L. y Gershuny, B. S. (1998). A structured interview for the assessment of the five-factor model of personality. *Psychological Assessment, 10,* 229–240.

Tuddenham, R. D., Blumenkrantz, J. y Wilkin, W. R. (1968). Age changes in AGCT: A longitudinal study of average adults. *Journal of Counseling & Clinical Psychology, 32,* 659–663.

Tyler, L. E. (1964). The antecedents of two varieties of interest pattern. *Genetic Psychology Monographs, 70,* 177–227.

Udai, P. (1995). Life-Orientation Inventory. En J. W. Pfeiffer (Ed.), *The 1995 annual: Vol. I, Training* (pp. 141–152). San Diego, CA: Pfeiffer & Co.

Underwood, B. y Moore, B. S. (1981). Sources of behavioral consistency. *Journal of Personality and Social Psychology, 40,* 780–785.

United States v. *City of Buffalo,* 37 U.S. 628 (W.D.N.Y. 1985).

United States v. *Georgia Power Company,* 5 FEP 587 (1973).

U.S. Department of Defense. (Septiembre de 1995). *ASVAB 18/19 Counselor Manual: The ASVAB career exploration program.* North Chicago, IL: HQ USMEPCOM/MOP-TA.

U.S. Department of Defense. (Diciembre 1999). *Technical manual for the ASVAB 18/19 career exploration program.* North Chicago, IL: HQ USMEPCOM.

U.S. Department of Education, Office for Civil Rights. (Julio 22, 1997). *1994 elementary and secondary school civil rights compliance report: Projected values for the nation.* Unpublished table.

U.S. Department of Labor, Employment and Training Administration, U.S. Employment Service. (1991, 1993). *Dictionary of occupational titles, 4th edition.* Washington, DC: Author.

U.S. Equal Employment Opportunity Commission. (1973, Ago. 23). *The uniform guidelines of employee selection procedures.* Discussion draft. Washington, DC: Author.

U.S. Equal Employment Opportunity Commission. (1994). *Enforcement guidance: Preemployment disability-related inquiries and medical examinations*

under the Americans with Disabilities Act (Notice Number 915.002). Washington, DC: Author.

U.S. Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. (1978). *Uniform guidelines on employee selection procedures.* 29 C.F.R. 1607.

Utz, P. y Korben, D. (1976). The construct validity of the occupational themes on the Strong–Campbell Inventory: *Journal of Vocational Behavior, 9,* 31–42.

Vale, C. D. (1985). *ASCAL: Item parameter estimation program (computer program).* St. Paul, MN: Assessment Systems, Inc.

Vernon, P. E. (1960). *The structure of human abilities* (rev. ed.). London: Methuen.

Vernon, P. E. (1979). Intelligence testing and the nature-/nurture debate, 1928–1978: What next? *British Journal of Educational Psychology, 49,* 1–14.

Vernon, P. E. (1985). Intelligence: Heredity–environment determinants. En T. Husén & T. N. Posthlethwaite (Eds.), *The international encyclopedia of education* (vol. 5, pp. 2605–2611). New York: Wiley.

Vispoel, W. P. (1999). Creating computerized adaptive tests of music aptitude: Problems, solutions, and future directions. En F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 151–176). Mahwah, NJ: Erlbaum.

Vollema, M. G. y Ormel, J. (2000). The reliability of the Structured Interview for Schizotypy-Revised. *Schizophrenia Bulletin, 26,* 619–629.

Voress, J. K. y Maddox, T. (1998). *Developmental Assessment of Young Children: Examiner's manual.* Austin, TX: pro.ed.

Wagner, M. E., Schubert, H. J. y Schubert, D. S. (1985). Family size effects: A review. *Journal of Genetic Psychology, 146,* 65–78.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer.* Mahwah, NJ: Erlbaum.

Wallach, M. A. y Kogan, N. (1965). *Modes of thinking in young children.* New York: Holt, Rinehart & Winston.

Waller, N. G. (1998a). Review of the Beck Anxiety Inventory. *Thirteenth Mental Measurements Yearbook,* 98–100.

Waller, N. G. (1998b). Review of the Beck Depression Inventory. *Thirteenth Mental Measurements Yearbook,* 120–121.

Waller, N. G., Lykken, D. T. y Tellegen, A. (1995). Occupational interests, leisure time interests, and personality: Three domains or one? Findings from the Minnesota Twin Registry. En D. J. Lubinski & R. V. Dawis (Eds.), *Assessing individual differences in human behavior: New concepts, methods, and findings* (pp. 233–259). Palo Alto, CA: Davies–Black Publishing/Consulting Psychologists Press.

Wallston, K. A. y Wallston, B. S. (1981). Health locus of control scales. En H. M. Lefcourt (Ed.), *Research with the locus of control construct* (Vol. 1, pp. 189–243). New York: Academic Press.

Wang, E. W., Rogers, R., Giles, C. L., Diamond, P. M., Herrington-Wang, L. E. y Taylor, E. R. (1997). A pilot study of the Personality Assessment Inventory (PAI) in corrections: Assessment of malingering, suicide risk, and aggression in male inmates. *Behavioral Sciences & the Law, 15,* 469–482.

Wards Cove Packing Company v. *Antonio et al.,* 490, U.S. 642 (1989).

Warnath, G. F. (1975). Vocational theories: Direction to nowhere. *Personnel and Guidance Journal, 53,* 422–428.

Washington v. *Davis,* 426 U.S. 229, 12 FEP 1415 (1976).

Watkins, C. E., Jr., Campbell, V. L. y Nieberding, R. (1994). The practice of vocational assessment by counseling psychologists. *Counseling Psychologist, 22,* 115–128.

Watkins, C. E., Jr., Campbell, V. L., Nieberding, R. y Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26,* 54–60.

Watkins, C. E., Jr., Campbell, V. L., Nieberding, R. y Hallmark, R. (1996). On Hunsley, harangue, and hoopla. *Professional Psychology: Research and Practice, 27,* 316–318.

Watson v. *Fort Worth Bank and Trust,* 487 U.S. 977, 108 S. Ct. 277 (1988).

Webb, E. (1915). Character and intelligence. *British Journal of Psychology Monograph Supplement,* III.

Webb, J. T. y Meckstroth, B. (1982). *Guiding the gifted child.* Columbus: Ohio Psychology Publishing Co.

Wechsler, D. (1981). *WAIS-R manual.* New York: Psychological Corporation.

Weinberg, R. A. (1989). Intelligence and IQ: Landmark issues and great debates. *American Psychologist, 44,* 98–104.

Weiner, I. B. (1983). The future of psychodiagnosis revisited. *Journal of Personality Assessment, 47,* 451–461.

Weiner, I. B. (1996). Some observations on the validity of the Rorschach Inkblot method. *Psychological Assessment, 8,* 206–213.

Weiss, J., Beckwith, B. y Schaeffer, B. (1989). *Standing up for the SAT.* New York: Simon & Schuster.

Wewers, M. E. y Lowe, N. K. (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing & Health, 13,* 227–236.

Wexley, K. N., & Klimoski, R. (1984). Performance appraisal: An update. In K. M. Rowland & G. R. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 2, pp. 35–79). Greenwich, CT: JAI Press.

White, L. J. (1996). Review of the Personality Assessment Inventory: A new psychological test for clinical and forensic assessment. *Australian Psychologist, 31,* 38–40.

White, N. y Cunningham, W. R. (1988). Is terminal drop pervasive or specific? *Journal of Gerontology: Psychological Sciences, 43,* P141–P144.

Whyte, W. H., Jr. (1956). *The organization man.* Garden City, NY: Doubleday.

Wiersma, U. y Latham, G. P. (1986). The practicality of behavioral observation scales, behavior expectation scales, and trait scales. *Personnel Psychology, 39,* 619–628.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment.* Reading, MA: Addison–Wesley.

Wiggins, N. y Kohen, E. S. (1971). Man versus model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology, 19,* 100–106.

Wilbur, P. H. (1970). Positional response set among high school students on multiple-choice tests. *Journal of Educational Measurement, 7,* 161–163.

Wildman, R., *et al*. (1980). *The Georgia Court Competency Test:* An attempt to develop a rapid, quantitative measure of fitness for trial. Unpublished manuscript, Forensic Services Division, Center State Hospital, Milledgeville, GA.

Willerman, L., Schultz, R., Rutledge, J. N. y Bigler, E. (1989). *Magnetic resonance imaged brain structures and intelligence.* Documento presentado en la 19a. reunión anual de Behavior Genetics Association. Charlottesville, VA.

Williams, W. M. y Ceci, S. J. (1997). Are Americans becoming more or less alike? Trends in race, class, and ability differences in intelligence. *American Psychologist, 52,* 1226–1235.

Willis, S. L. (1990). Introduction to the special section on cognitive training in later adulthood. *Developmental Psychology, 26,* 875–878.

Willson, V. L. (1982). Maximizing reliability in multiple-choice questions. *Educational & Psychological Measurement, 42,* 69–72.

Wilson, R. S. (1983). The Louisville Twin Study: Developmental synchronies in behavior. *Child Development, 54,* 298–316.

Wilson, R. S. (1985). Risk and resilience in early mental development. *Developmental Psychology, 21,* 795–805.

Wingard, J. A. y Maltzman, I. (1980). Interest as a predeterminer of the GSR index of the orienting reflex. *Acta Psychologica, 46,* 153–160.

Winner, E. (1996). *Gifted children: Myths and realities.* New York: Basic Books.

Witelson, S. F., Glezer, I. I. y Kigar, D. L. (1995). Women have greater density of neurons in posterior temporal cortex. *Journal of Neuroscience, 15,* 3418– 3428.

Wober, M. (1974). Towards an understanding of the Uganda concept of intelligence. En J. W. Berry & P. R. Dasen (Eds.), *Culture and cognition: Readings in cross-cultural psychology.* London: Methuen.

Wolff, W. T. y Merrens, M. R. (1974). Behavioral assessment: A review of clinical methods. *Journal of Personality Assessment, 38,* 3–16.

Wolins, L. y Dickinson, T. T. (1973). Transformations to improve reliability and/or validity for affective scales. *Educational & Psychological Measurement, 33,* 711–713.

Wolk, R. y Wolk, R. (1971). *The Gerontological Apperception Test.* New York: Behavioral Publications.

Woodcock, R. W. (1998). Extending Gf–Gc theory into practice. En J. J. McArdle & R. W. Woodcock (Eds.), *Human cognitive abilities in theory and practice* (pp. 137–156). Mahwah, NJ: Erlbaum.

Woodmansee, J. J. (1970). The pupil response as a measure of social attitude. En G. F. Summers (Ed.), *Attitude measurement* (pp. 514–534). Chicago: Rand McNally.

Woodrum, E. y Ventis, W. L. (1992). Moral Attitudes Index. *Journal of Empirical Theology, 5,* 70–84.

World Health Organization. (1992). *International statistical classification of diseases and related health problems* (10a. revisión, ICD-10). Geneva: Author.

Wright, B. D., & Linacre, M. (1991). *A user's guide to BIGSTEPS.* Chicago: MESA Press.

Wright, H. E. (1960). Observational child study. En P. E. Mussen (Ed.), *Handbook of research methods in child development* (pp. 71–139). New York: Wiley.

Wrightsman, L. S. (1994). *Adult personality development. Vol. 1. Theories and concepts.* Thousand Oaks, CA: Sage.

Yang, S.-Y. y Sternberg, R. J. (1997). Taiwanese Chinese people's conceptions of intelligence. *Intelligence, 25,* 21–36.

Yerkes, R. M. (Ed.). (1921). Psychological examining in the United States army. *Memoirs of the National Academy of Sciences,* vol. 15.

Zajonc, R. B. (1986). The decline and rise of scholastic aptitude scores. *American Psychologist, 41,* 862–867.

Zarske, J. A. (1985). Review of Adjective Check List. *Ninth Mental Measurements Yearbook,* 52–53.

Zeskind, P. S. y Ramey, C. T. (1981). Preventing intellectual and interactional sequelae of fetal malnutrition: A longitudinal transaction and synergistic approach to development. *Child Development, 52,* 213–218.

Zigler, E. (1988). The IQ pendulum. [Review of the book by H. H. Spitz, The raising of intelligence: A selected

history of attempts to raise retarded intelligence.] *Readings: A Journal of Reviews and Commentary in Mental Health, 3,* 4–9.

Zigler, E. y Hodapp, R. M. (1986). *Understanding mental retardation.* New York: Cambridge University Press.

Ziskin, J. (1986). The future of clinical assessment. En B. S. Plake & J. C. Wirt (Eds.), *The future of testing* (pp. 185–201). Hillsdale, NJ: Erlbaum.

Zook, J. (1993). Two agencies start work on national test of college students' analytical skills. *Chronicle of Higher Education, 39*(29), A23.

Zuckerman, M. y Lubin, B. (1985). *Manual for the Multiple Affect Adjective Check List–Revised.* San Diego, CA: EdITS.

Zytowski, D. G. (1976). Predictive validity of the Kuder Occupational Interest Survey: A 12- to 19-year follow-up. *Journal of Counseling Psychology, 23,* 221–233.

ÍNDICE DE AUTORES

ÍNDICE DE MATERIAS

ÍNDICE DE TESTS