Tests Psicológicos

Datos de catalogación bibliográfica

ANASTASI ANNE, URBINA SUSANA
Tests Psicológicos
PRENTICE HALL, México, 1998

ISBN: 970-17-0186-0
Materia: Psicología

Formato: 19 x 23.5 Páginas: 744

EDICIÓN EN ESPAÑOL:

GERENTE EDITORIAL: LUIS GERARDO CEDEÑO PLASCENCIA
SUPERVISOR DE TRADUCCIÓN: JOSÉ LUIS NÚÑEZ HERREJÓN
SUPERVISOR DE EDICIÓN: JOSÉ D. HERNÁNDEZ GARDUÑO
CORRECTOR DE ESTILO: JOSÉ FRANCISCO JAVIER DÁVILA MARTÍNEZ

Edición en inglés:

Editor in Chief: Pete Janzow
Acquisitions Editor: Heidi Freund
Director of Production and Manufacturing: Barbara Kittle
Managing Editor: Bonnie Biller
Manufacturing Manager: Nick Sklitsis
Prepress and Manufacturing Buyer: Tricia Kenny
Creative Design Director: Leslie Osher
Marketing Manager: Michael Alread
Art Coordinator: Michele Giusti
Editorial Assistant: Emsal Hasan
Cover Design/Interior Design: Circa '86
Cover Photo Credit: Kasimir Malevich, *Suprematism*, c.1917, oil on canvas, 80 × 80 cm., Museum of Fine Arts, Krassnodar, Russia, Erich Lessing / Art Resources, NY
Acknowledgments begin on page 681, which constitutes a continuation of the copyright page.

ANASTASI: TESTS PSICOLOGICOS, 7a. Ed.

Traducido de la séptima edición en inglés de la obra: **Pshychological Testing.**

PRENTICE-HALL HISPANOAMERICANA, S.A.

Atlacomulco Núm. 500-5° Piso
Col. Industrial Atoto
53519, Naucalpan de Juárez, Edo. de México

ISBN 970-17-0186-0
Miembro de la Cámara Nacional de la Industria Editorial, Reg. Núm. 1524.
Original English Language Edition Published by Prentice-Hall, Inc. A Simon & Schuster Company

ISBN 0-02-303085-2

IMPRESO EN MÉXICO/PRINTED IN MEXICO

Contenido

Prefacio

La década de los noventa ha presenciado la continuación y el crecimiento del renovado interés en las pruebas psicológicas que se hizo evidente en los años ochenta, como lo muestra el desarrollo de recientes pruebas —algunas de las cuales representan nuevos planteamientos— así como la revisión y la continuada investigación de las anteriores. Al elegir los instrumentos que queríamos citar o analizar, la primera meta era exponer al lector a la rica variedad de herramientas de medición de que se dispone actualmente en el campo, lo mismo que a algunas pruebas y técnicas de importancia histórica; cualquier tentativa por cubrir exhaustivamente el campo, o incluso una parte significativa, es una tarea que desde luego rebasa el alcance de este libro.

Cada vez es mayor la atención que se concede al sujeto examinado, por lo que se exhorta al usuario de las pruebas a investigar, en el caudal de experiencias y reacciones propias, las causas de su desempeño: ¿Qué hay en los antecedentes del individuo que nos ayude a entender sus respuestas a la prueba?, ¿cuáles de sus elementos pueden aumentar el valor de los resultados para predecir su conducta ulterior en la escuela, el trabajo y otras actividades cotidianas? De lo anterior se desprende que ha aumentado la responsabilidad del *usuario de la prueba* en cuanto a la selección de los instrumentos adecuados, los métodos para presentarlos al individuo, la interpretación de los resultados y la comunicación y utilización de los mismos. Debido a dichas consideraciones, la obra está especialmente diseñada para brindar una base que permita el uso adecuado de las pruebas.

El empleo eficaz de los instrumentos requiere de cierta familiaridad con su elaboración, pues esta información es necesaria para evaluar distintas pruebas, elegir la apro-

piada para determinado propósito o sujeto e interpretar adecuadamente las puntuaciones. Aunque este libro no se dirige en particular a quienes elaboran los instrumentos, incluye información suficiente sobre su realización para satisfacer las necesidades del usuario.

Esta edición también proporciona explicaciones sencillas de algunos conceptos y procedimientos de uso común y rápido desarrollo que tal vez influyan en las prácticas psicométricas del siglo XXI, como es el caso de los tests adaptados a las computadoras (*computerized adaptive testing*, CAT), el metanálisis, el modelamiento de ecuaciones estructurales, el uso de intervalos de confianza en lugar de la significancia estadística tradicional, las pruebas transculturales y el uso creciente del análisis factorial en el desarrollo de los tests de habilidad y de personalidad. Una aplicación práctica del método de análisis de factores ha sido el establecimiento de normas que permitan la interpretación de las puntuaciones en diferentes niveles de especificidad o generalidad, de modo que el usuario aplique el más adecuado a la persona o situación particular.

En el uso actual de las pruebas se han hecho muy evidentes dos tendencias de gran importancia a largo plazo; en lugar de dedicarles capítulos aparte, decidimos analizarlas siempre que resultó conveniente en todo el libro. La primera es la creciente influencia de la computación en el desarrollo, la elaboración y aplicación de los instrumentos, además de los usos conocidos de las computadoras en la calificación de las pruebas y el procesamiento de las puntuaciones. La velocidad de los avances tecnológicos es tan sorprendente que parece marcar el ritmo del progreso en áreas importantes de la psicología. Asimismo, la tecnología también hace contribuciones notables a los frentes teóricos y de investigación; por ejemplo, en la bibliografía psicológica se observa una gran cantidad de integración y fecundación de distintas áreas, situación que se ve favorecida por la facilidad con la que los investigadores de todo el mundo tienen acceso a la información, la procesan y la comunican. Uno de los casos más destacados y prometedores de esta tendencia a la integración es el replanteamiento de los rasgos cognoscitivos y de personalidad como aspectos inseparables del individuo, que a su vez está inextricablemente ligado a su yo físico, a los sucesos de su vida y a su ambiente.

La segunda tendencia significativa en las pruebas psicológicas es la creciente intrusión de intereses políticos y legales. Aunque se trata de una influencia que provoca divisiones y, por ende, es potencialmente nociva, también puede tener ramificaciones positivas pues fomenta la creatividad y aumenta la vigilancia de las consecuencias, deliberadas o no, del uso de las pruebas. A lo largo del texto se citan algunas leyes promulgadas en los Estados Unidos que han tenido un gran efecto en las prácticas psicométricas, incluidos el título y el año en que fueron aprobadas; su contenido puede encontrarse en el *Congressional Record* de los Estados Unidos, así como en otras publicaciones periódicas en las secciones de referencia de la mayor parte de las bibliotecas de ese país.

Las primeras seis ediciones de este libro fueron escritas por una sola persona; en cambio, ésta es producto de una verdadera colaboración. Las dos autoras planearon la reorganización de los capítulos y los temas principales. La revisión y reelaboración de los capítulos se distribuyó de la siguiente manera: Anastasi fue responsable de los capítulos 1 a 7 y 10 a 12, mientras que Urbina se encargó de los capítulos 8, 9 y del 13 al 18 y asumió mayores funciones administrativas y de correspondencia. Además, cada

una leyó el borrador de los capítulos de la otra y sugirió mejoras que en general fueron aceptadas.

Obviamente, este libro no se hubiera escrito sin el acceso a las investigaciones y las publicaciones de muchos psicólogos de diversas partes del mundo, tanto contemporáneos como de épocas pasadas. Sus nombres aparecen en las citas de sus publicaciones y en la lista bibliográfica que se encuentra al final de la obra. Dentro de este grupo impresionante destacan algunos individuos por su puntual cooperación y la magnitud de sus contribuciones. Entre ellos sobresalen Dianne Brown del Directorio Científico de la Asociación Psicológica Estadounidense (*American Psychological Association, APA*), Aurelio Prifitera y Joanne Lenke de la *Psychological Corporation*, Lorin Letendre de *Consulting Psychologists Press*, Carol Watson de *National Computer Systems*, Douglas Jackson de *Sigma Assessment Systems*, Elizabeth McGrath y John Oswald de la *Riverside Publishing Company* y Wayne Camara del *College Board*. Además, reconocemos con gratitud la ayuda que nos prestaron los miembros del personal de las bibliotecas de las universidades Fordham y del Norte de Florida para enfrentar los retos que presentó la preparación de esta obra.

A.A.
S.U.

Naturaleza y uso de las pruebas psicológicas

Las pruebas, tests o instrumentos de medición psicológicos son herramientas, y para obtener los beneficios que proporcionan es necesario tener presente este hecho esencial. Cualquier herramienta puede ser un medio para hacer el bien o el mal, dependiendo de cómo se emplee. Las pruebas se han desarrollado a un paso creciente, y aunque cada vez son más las áreas de la vida cotidiana a las que contribuyen,[1] este crecimiento ha estado acompañado de algunos abusos y de expectativas poco realistas. El usuario de los tests necesita saber cómo evaluarlos. ¿Qué tan buena es esta prueba para el propósito que se pretende que cumpla? ¿Qué información puede brindar sobre la persona a la que se aplica? ¿Cómo pueden integrarse sus resultados en la red de datos que se utiliza en la toma de decisiones? Escribimos este libro desde el punto de vista de estas preguntas, por lo que no se dirige al especialista, sino más bien al estudiante de psicología. En la actualidad se requiere de ciertos conocimientos básicos sobre los instrumentos de medición no sólo entre quienes los elaboran o aplican, sino también de parte de cualquiera que se sirva de sus resultados como fuente de datos para tomar decisiones acerca de sí mismo o de los demás.

USOS Y VARIEDADES DE LAS PRUEBAS PSICOLÓGICAS

Habitualmente, la función de las pruebas psicológicas ha sido medir las diferencias entre individuos o entre las reacciones de la misma persona en circunstancias distintas. El diagnóstico del retardo mental fue uno de los primeros problemas que esti-

[1]Véase Dahlstrom (1993b) para una lúcida ilustración de las contribuciones de las pruebas psicológicas con ejemplos reales.

mularon su desarrollo, y, hasta el momento, la detección de las deficiencias intelectuales sigue siendo una aplicación importante de ciertos instrumentos. Los usos clínicos incluyen el examen de personas con trastornos emocionales graves y otros problemas de conducta. La evaluación de las necesidades educativas dio un fuerte impulso al desarrollo inicial de las pruebas, como fue el caso de los famosos tests de Binet que comenzaron el movimiento de las mediciones de la inteligencia. En la actualidad, las escuelas se cuentan entre los principales usuarios, ya que les permiten, entre otras muchas cosas, clasificar a los niños según su capacidad para beneficiarse de las diferentes formas de instrucción escolar, identificar a los excepcionalmente lentos o rápidos para aprender, brindar asesoría educativa y vocacional a los estudiantes de educación media y superior, y seleccionar a los aspirantes a las escuelas profesionales.

La selección y clasificación del personal industrial es otra aplicación fundamental de las pruebas psicológicas. Desde el operador de la línea de montaje y el archivista, hasta las funciones directivas, difícilmente puede encontrarse un puesto para el que alguna prueba no haya demostrado su utilidad psicológica en cuanto a contratación, asignación de tareas, transferencias, ascensos o despidos. En muchas de estas situaciones, en especial cuando se relacionan con los puestos de nivel superior, se requiere que las pruebas se empleen junto con una entrevista realizada por un experto que, al interpretar las puntuaciones a la luz de otra información importante sobre el individuo, las aprovecha mejor. Con todo, la aplicación de pruebas constituye una parte importante del programa global del departamento de personal. Una aplicación relacionada de las pruebas psicológicas se encuentra en la selección y clasificación del personal militar. Luego de sus inicios en la Primera Guerra Mundial, el alcance y la variedad de los instrumentos de medición psicológicos usados en contextos militares mostraron un desarrollo notable durante la Segunda Guerra Mundial. Posteriormente, su investigación y desarrollo ha continuado a gran escala y en todas las ramas de las fuerzas armadas.

En la consejería individual el uso de pruebas ha aumentado gradualmente de una orientación limitada a los planes educativos y vocacionales al interés en todos los aspectos de la vida de la persona. El bienestar emocional y las relaciones personales adecuadas se han convertido en objetivos prominentes de la consejería. También se observa una tendencia a servirse de las pruebas para aumentar el desarrollo y la comprensión personales. En este marco, las puntuaciones de los instrumentos son parte de la información que se proporciona al individuo para ayudarlo a tomar decisiones.

Resulta evidente el uso de los instrumentos de medición psicológicos en la solución de una gran variedad de problemas prácticos; sin embargo, no hay que perder de vista el hecho de que las pruebas también cumplen funciones importantes en la investigación básica. Por ejemplo, en casi todos los problemas de psicología diferencial se aplican tests para obtener datos, como ocurre con los estudios sobre la naturaleza y el grado de las diferencias individuales, la organización de los rasgos psicológicos, la medición de las diferencias grupales y la identificación de los factores biológicos y culturales asociados con las variaciones conductuales. En todas estas áreas de investigación —como en muchas otras— es fundamental la medición precisa de las diferencias individuales que las pruebas bien formuladas hacen posible. Del mismo modo, las pruebas

psicológicas proporcionan herramientas estandarizadas para investigar problemas tan diversos como los cambios que sufre el individuo a lo largo del ciclo de desarrollo, la eficacia relativa de distintos procedimientos educativos, los resultados de la psicoterapia, el impacto de los programas comunitarios y la influencia de las variables ambientales en el desempeño.

Las pruebas diseñadas para estos diversos propósitos también difieren en otras características notables. Varían en la forma en que se aplican, ya sea que el examinador capacitado trabaje con un individuo y luego con otro o de manera simultánea con grupos grandes, o bien por medio de una computadora. También difieren en los aspectos de la conducta que cubren. Algunas se concentran en la evaluación de los rasgos cognoscitivos o las habilidades, que pueden ir de aptitudes generales —como la capacidad de beneficiarse de la educación universitaria— a las habilidades sensoriomotoras muy especializadas que se requieren para realizar una operación manual sencilla. Otros instrumentos miden las variables afectivas o de personalidad, como los rasgos emocionales o motivacionales, la conducta interpersonal, los intereses, las aptitudes y los valores.

Frente a tal diversidad de naturaleza y propósito, ¿qué características tienen en común las pruebas psicológicas?, ¿en qué se diferencian de otros métodos para obtener información acerca de los individuos? La respuesta se encuentra en ciertos rasgos fundamentales de su elaboración y uso, que constituyen el punto de interés de este capítulo.

¿QUÉ ES UNA PRUEBA PSICOLÓGICA?

Una muestra de conducta. En esencia, la prueba psicológica es una medida objetiva y estandarizada de una muestra de conducta. Con las pruebas psicológicas, como con las de cualquier otra ciencia, se hacen observaciones sobre una *muestra* pequeña, pero cuidadosamente elegida, de la conducta del individuo. A este respecto, el psicólogo procede de la misma manera que el bioquímico que analiza la sangre de un paciente o el suministro de agua de la comunidad, examinando una o más muestras. Si el psicólogo desea probar el léxico de un niño, la habilidad de un escolar para realizar cálculos aritméticos o la coordinación visomotora de un piloto, prueba su desempeño en un conjunto representativo de palabras, problemas aritméticos o pruebas motrices. Que el instrumento cubra adecuadamente o no la conducta considerada depende obviamente del número y la naturaleza de los reactivos de la muestra; por ejemplo, una prueba de aritmética que no tenga más que cinco problemas o que sólo incluya multiplicaciones sería una mala medida de la habilidad del sujeto para hacer operaciones. Una prueba de vocabulario compuesta en exclusiva por términos provenientes de la jerga del béisbol difícilmente podría brindar una estimación confiable del léxico del niño.

El *valor de diagnóstico* o *predictivo* de un test psicológico depende de qué tanto funcione como indicador de un área de conducta relativamente amplia y significativa. La medición de la muestra de conducta que examina la prueba rara vez, si acaso, es el objetivo. El conocimiento del niño de una lista particular de 50 palabras no es, en sí, de gran interés, como tampoco es de mayor importancia el desempeño de quien soli-

cita un empleo en un conjunto de 20 problemas aritméticos. Pero los tests cumplen su propósito si muestran que hay una correspondencia estrecha entre el conocimiento que el niño tiene de la lista de palabras y su dominio del vocabulario o entre la puntuación que obtiene el solicitante en los problemas aritméticos y su desempeño en el empleo.

A este respecto, hay que observar que no es necesario que los reactivos se asemejen a la conducta que la prueba pretende predecir, lo único que se requiere es demostrar una correspondencia empírica entre ambos, de ahí que el grado de similitud entre la muestra de la prueba y la conducta por predecir varíe mucho. En un extremo, la prueba puede coincidir por completo con una parte de la conducta que se quiere predecir, como en el caso de una prueba de vocabulario de una lengua extranjera que examina al estudiante en 20 de las 50 palabras estudiadas o el de la prueba de conducción para obtener la licencia de manejo. Un grado menor de semejanza se encuentra en muchas pruebas de aptitud vocacional que se aplican antes de la capacitación para el trabajo, en las que hay apenas un parecido moderado entre las tareas que se realizan en el puesto y las que incluye la prueba. En el otro extremo se encuentran los tests proyectivos de personalidad, como el de manchas de tinta de Rorschach, en el que a partir de las asociaciones que el examinado hace de las manchas se intenta predecir su reacción a otras personas, a estímulos emocionales y a otras situaciones complejas de la vida cotidiana. A pesar de sus diferencias superficiales, todas esas pruebas constan de muestras de conducta del individuo, y cada una debe probar su valor con la demostración de una correspondencia empírica entre el desempeño del examinado en la prueba y en otras situaciones.

También representa una distinción menor que se utilice el término "diagnóstico" o "predicción". Por lo general, el segundo connota una estimación temporal; por ejemplo, se pronostica el desempeño del individuo en un trabajo a partir de su ejecución en la prueba. Pero en un sentido más amplio incluso el diagnóstico de una condición actual, como el retardo mental o un trastorno emocional, lleva implícita la predicción de lo que el individuo hará en otras situaciones. Lógicamente, es más sencillo considerar todas las pruebas como muestras de conducta a partir de las cuales se hacen predicciones que atañen a otro comportamiento. Entonces, es posible caracterizar las diferentes clases de pruebas o tests como variaciones de este patrón básico.

Otro punto que debemos considerar desde el principio tiene que ver con el concepto de *capacidad*. Por ejemplo, es totalmente posible elaborar una prueba para predecir qué tan bien aprenderá el francés un alumno antes de que empiece el curso. La prueba debería incluir una muestra de los comportamientos que se requieren para aprender el nuevo idioma y también presuponer que el estudiante no tiene ningún conocimiento. En este caso podría decirse que la prueba mide la "capacidad" o "potencialidad" del individuo para aprender el francés. Sin embargo, hay que tener cautela al emplear esos términos en relación con las pruebas psicológicas. Únicamente podemos decir que una prueba mide la "capacidad" en el sentido de que una muestra de la conducta actual puede utilizarse como indicador de otra conducta futura. Ninguna prueba psicológica puede hacer más que medir el comportamiento, y que éste sirva como índice de otra conducta sólo lo establece un experimento empírico.

Estandarización. Recordemos que, en la definición inicial, dijimos que la prueba psicológica es una medida estandarizada. La estandarización supone la *uniformidad de los procedimientos* en la aplicación y calificación de la prueba. Es evidente que si los resultados que obtienen distintas personas han de ser comparables, las condiciones del examen tienen que ser las mismas para todos. Tal requisito es sólo una manifestación de la necesidad de tener condiciones controladas en todas las observaciones científicas. En una situación de prueba, la única variable independiente es a menudo el individuo examinado.

Para asegurar la uniformidad de las condiciones de prueba, quien la elabora proporciona instrucciones detalladas para la aplicación de cada nuevo instrumento. La formulación de las instrucciones es una parte importante de la estandarización de la nueva prueba, y se extiende a los materiales exactos que debe emplearse, los límites de tiempo, las instrucciones orales, las demostraciones previas, las formas de manejar las dudas de los examinados y cualquier otro detalle de la situación de examinación. Así, al dar instrucciones o presentar oralmente los problemas, hay que considerar la velocidad con que se habla, el tono de la voz, la inflexión, las pausas y la expresión del rostro; por ejemplo, en una prueba que consiste en detectar absurdos, uno puede dar la respuesta correcta al sonreír o al hacer una pausa cuando se lee la palabra crucial. En una sección posterior del capítulo, que trata de los problemas de la aplicación de las pruebas, veremos el procedimiento de estandarización.

Otro paso importante en la estandarización de las pruebas es el establecimiento de *normas*. Las pruebas psicológicas no tienen criterios predeterminados de aprobación o reprobación; el desempeño en cada prueba se evalúa sobre la base de los datos empíricos. Para la mayor parte de los propósitos, a fin de interpretar el resultado que obtiene el individuo en una prueba, éste se compara con los resultados de otros en la misma prueba. Como lo sugiere el término, la norma es el desempeño normal o promedio. De esta manera, si en una prueba de razonamiento aritmético los niños normales de ocho años resuelven correctamente 12 de 50 problemas, entonces, en esta prueba, la norma para los ocho años corresponde a una puntuación de 12. Esto se conoce como puntuación cruda (bruta o directa), y se expresa como el número de reactivos correctos, el tiempo requerido para completar la tarea, el número de errores o alguna otra medida objetiva que sea adecuada para el contenido de la prueba. Esta puntuación cruda no tiene sentido hasta que no se evalúa en términos de datos interpretativos adecuados.

Durante el proceso de estandarización, la prueba se aplica a una muestra grande y representativa de las personas a las que va dirigida. Este grupo, conocido como muestra de estandarización, sirve para establecer las normas, que indican no sólo el desempeño promedio sino también la frecuencia relativa de las desviaciones por encima y por debajo del promedio, lo que permite evaluar diferentes grados de superioridad e inferioridad. En el capítulo 3 consideraremos las formas concretas de expresar tales normas, que permiten designar la posición del individuo en relación con la muestra normativa o de estandarización.

También conviene observar que, para los tests de personalidad, las normas se establecen esencialmente de la misma manera que para los de aptitud. En uno de personalidad la norma no es por fuerza la ejecución más deseable o "ideal", como tampoco

una puntuación perfecta o sin errores es la norma de un test de aptitud. En ambas pruebas la norma corresponde a la ejecución de la persona promedio. Por ejemplo, en las pruebas de dominancia-sumisión, la norma cae en un punto medio que representa el grado de dominio o de sumisión que manifiesta la persona promedio. De modo similar, en un inventario de ajuste emocional la norma por lo general no corresponde a una ausencia absoluta de respuestas inadaptadas o desfavorables. La mayoría de los individuos "normales" de la muestra de estandarización presenta algunas de esas respuestas, y este número de respuestas inadaptadas, por consecuencia, podría representar la norma.

Medición objetiva de la dificultad. Al iniciar este análisis definimos las pruebas psicológicas como una medición objetiva y estandarizada. ¿En qué sentido específico decimos que es objetiva? Ya tocamos algunos aspectos de la objetividad al hablar de la estandarización. En efecto, la aplicación, calificación e interpretación de los resultados serán objetivas en la medida en que sean independientes del juicio subjetivo del examinador. En teoría, cualquier individuo al que se aplique la prueba puede obtener una puntuación idéntica independientemente de quién la aplique. Por supuesto, esto no es del todo cierto porque en la práctica no se han alcanzado la estandarización ni la objetividad perfectas. Pero al menos la objetividad es la meta de la elaboración de instrumentos y casi todos la demuestran en un grado razonablemente elevado.

Hay otras condiciones que permiten señalar a las pruebas psicológicas como objetivas. La determinación del grado de dificultad de un reactivo o de toda la prueba se basa en procedimientos objetivos empíricos. Cuando Binet y Simon prepararon en 1905 su escala original para la medición de la inteligencia, distribuyeron los 30 reactivos de la escala en orden de dificultad creciente, que determinaron luego de probar los reactivos en 50 niños normales y en algunos con retardo mental. Tomaron los reactivos que resolvió correctamente el mayor número de niños, *ipso facto*, como los más sencillos, mientras que consideraron más difíciles los que pocos solucionaron. Con este procedimiento, establecieron un orden empírico de dificultad. Este primer ejemplo es característico de la medición objetiva del nivel de dificultad, que ahora es una práctica común en la elaboración de pruebas psicológicas.

No sólo el ordenamiento, sino también la selección de reactivos para su inclusión en una prueba, puede apoyarse en la proporción de sujetos de la muestra que resuelve cada reactivo. Así, si hay muchos reactivos en el extremo sencillo o el difícil de la escala, es posible descartar algunos. De modo similar, si en ciertas partes de la graduación de la dificultad de los reactivos son escasos, es posible agregar otros para llenar las lagunas. En el capítulo 7 trataremos aspectos más técnicos del análisis de reactivos.

Confiabilidad. ¿Qué tan buena es la prueba? ¿En realidad funciona? Estas preguntas podrían —y ocasionalmente lo hacen— exigir largas horas de análisis infructuoso. Por una parte, las opiniones subjetivas, las corazonadas y los sesgos personales pueden conducir a afirmaciones extravagantes respecto de lo que puede lograr una prueba o, por la otra, a un rechazo obstinado. La única forma de que estas preguntas reciban una respuesta concluyente es realizar una comprobación empírica. La *evaluación objetiva*

de las pruebas psicológicas consiste principalmente en determinar su confiabilidad y validez en situaciones especificadas.

Como se emplea en la psicometría, el término "confiabilidad" significa básicamente consistencia. La confiabilidad de una prueba es la consistencia de las puntuaciones obtenidas por las mismas personas cuando se les aplica la misma prueba o una forma equivalente. Si un niño tiene un CI de 110 el lunes y uno de 80 el viernes, es obvio que no se puede confiar mucho en ninguna de las dos puntuaciones. Asimismo, si, de un grupo de 50 palabras, el examinado identifica bien 40, mientras que, de otro grupo supuestamente equivalente, obtiene una puntuación de 20 correctas, ninguna de las puntuaciones puede considerarse como un indicador confiable de su comprensión verbal. Es posible que en ambos ejemplos sólo una de las puntuaciones sea errónea, pero esto sólo se demuestra con exámenes posteriores. De los datos obtenidos, lo único que se concluye es que no pueden ser correctos ambos, y sin información adicional es imposible establecer que uno o ninguno sea una estimación adecuada de la habilidad del individuo.

Antes de permitir la libre circulación de una prueba psicológica debe llevarse a cabo una verificación cuidadosa y objetiva de su confiabilidad. En el capítulo 4 estudiaremos los distintos tipos de confiabilidad, así como los métodos de medición de cada uno. Para comprobar la confiabilidad se comparan las puntuaciones obtenidas por las mismas personas en diversos momentos con diferentes conjuntos de reactivos, examinadores o calificadores, o en cualquier otra condición de examinación pertinente. Es esencial especificar el tipo de confiabilidad y el método empleado para determinarlo, ya que la misma prueba puede variar en esos diferentes aspectos. También hay que informar del número y la clase de personas con las que se hizo la verificación. Con estos datos, los usuarios pueden predecir si la prueba será tan confiable para el grupo al que esperan aplicarla, o si es probable que sea mayor o menor.

Validez. Es indudable que la pregunta más importante sobre cualquier prueba psicológica atañe a su validez —es decir, el grado con el que verdaderamente mide lo que pretende medir—. La validez proporciona una comprobación directa de qué tan bien cumple una prueba su función. Por lo general, para determinarla se requiere de *criterios* independientes y externos de lo que la prueba intenta medir. Por ejemplo, si se quiere emplear una prueba de aptitud médica para seleccionar, entre los aspirantes para ingresar a la escuela de medicina, a los más prometedores, un criterio puede ser el éxito de los seleccionados en la escuela. Durante el proceso de validación, la prueba debe aplicarse a un grupo grande de estudiantes en el momento de su admisión. Posteriormente tiene que obtenerse una medida del desempeño académico de cada uno sobre la base de sus notas, la calificación que reciban de los profesores, la terminación o el abandono de sus estudios y cosas similares. Esta medida constituye el criterio con el que se correlaciona la puntuación que recibió al principio cada estudiante. Una correlación, o *coeficiente de validez*, elevada significa que los individuos que en la prueba obtuvieron una calificación relativamente alta han sido más o menos exitosos en la escuela de medicina, mientras que los que obtuvieron bajas calificaciones en la prueba han tenido un pobre desempeño académico. Una correlación baja indicaría que existe poca correspondencia entre la puntuación

obtenida en la prueba y la medida considerada como criterio y, por lo tanto, que la validez de la prueba es poca. El coeficiente de validez nos permite determinar qué tan bien se predice el desempeño que se toma como criterio a partir de las puntuaciones de la prueba.

Las pruebas diseñadas para otros propósitos se validan de manera similar contra criterios apropiados. Por ejemplo, una prueba de aptitud vocacional puede validarse con el éxito laboral de un grupo experimental de nuevos empleados; una batería de aptitud para pilotos, con los resultados en los vuelos de entrenamiento. Las pruebas destinadas a usos más amplios y variados se validan con una serie de indicadores conductuales obtenidos de modo independiente, y su validez sólo puede establecerse con la acumulación gradual de datos de muchas investigaciones diferentes

Tal vez el lector haya notado una paradoja aparente en el concepto de validez de la prueba. Si es necesario hacer un seguimiento de los examinados u obtener de otro modo medidas independientes de lo que la prueba pretende medir, ¿por qué no prescindir de la prueba? La respuesta a este acertijo se encuentra en la distinción entre el grupo de validación por un lado y, por otro, los grupos a los que se aplicará la prueba con propósitos operativos. Antes de que la prueba esté lista para su uso es necesario establecer su validez con una muestra representativa de personas cuyas calificaciones no se emplean con propósitos operativos, sino sólo en el proceso de comprobación del instrumento. Si la prueba demuestra ser válida con ese método, puede utilizarse con otras muestras en ausencia de las medidas de criterio.

Aún podría argumentarse que sólo se necesita esperar a que la medida de criterio madure —que esté disponible— en *cualquier* grupo para obtener la información que la prueba trata de predecir. Pero semejante procedimiento supone un desperdicio tal de tiempo y energía que resultaría prohibitivo en casi todos los casos. Así, para determinar qué solicitantes tienen éxito en un empleo o qué estudiantes terminan con éxito la universidad, admitiríamos a todo aquel que lo solicite (o a una muestra aleatoria) y esperaríamos a ver que pasé. Las pruebas están diseñadas para disminuir al mínimo el derroche que supone este procedimiento, así como su nocivo impacto emocional en los individuos. Por medio de las pruebas es posible evaluar, con un margen de error determinable, el nivel actual de la persona en las habilidades requeridas, sus conocimientos así como otras características pertinentes. Entre mayor sea la validez y la confiabilidad de la prueba, menor será el margen de error.

En los capítulos 5 y 6, estudiaremos tanto los problemas especiales que uno enfrenta al determinar la validez de diversas pruebas como los criterios y los procedimientos estadísticos utilizados; sin embargo, en este momento es necesario considerar otro punto. La validez no sólo nos indica el grado en que la prueba cumple con su función, pues al estudiar los datos de la validación podemos determinar con objetividad *qué* es lo que mide el instrumento. En consecuencia, sería más preciso definir la validez como el grado en que sabemos qué es lo que mide la prueba. La interpretación de las puntuaciones sería indudablemente más clara y menos ambigua si las pruebas recibieran su nombre de acuerdo con las relaciones empíricas que las validaron. Se observa una tendencia en esta dirección en nombres como "prueba de evaluación académica" y "prueba de clasificación de personal" en lugar de títulos tan vagos como "test de inteligencia".

¿POR QUÉ CONTROLAR EL USO DE LAS PRUEBAS PSICOLÓGICAS?

"¿Puede venderme un ejemplar del Stanford-Binet? La próxima semana mi sobrino debe presentar una de las pruebas para ser admitido en la escuela X y necesita practicar para poder pasarla."

"Para mejorar el programa escolar de lectura necesitamos una prueba de CI justa que mida el potencial innato de cada niño."

"Anoche contesté las preguntas de un test de inteligencia publicado en una revista y obtuve un CI de 80. Me parece que las pruebas psicológicas no tienen sentido."

"Mi compañera de cuarto, que estudia psicología, me aplicó un test de personalidad y resulté neurótica. Desde entonces me he sentido muy molesta como para ir a clases."

"El año pasado, mientras realizaba una investigación usted aplicó a nuestros empleados un nuevo test de personalidad. Quisiéramos tener los resultados en sus expedientes."

Estos comentarios no son imaginarios; se basan en incidentes reales y cualquier psicólogo podría ampliar la lista. Ilustran abusos o malas interpretaciones de los instrumentos de medición psicológicos que podrían restarles todo valor o lastimar al individuo. Como cualquier instrumento científico o herramienta de precisión, las pruebas deben utilizarse correctamente para que sean eficaces. En manos de un usuario poco escrupuloso o bien intencionado pero ignorante pueden causar un grave daño. Hay dos razones principales para controlar su uso: (a) garantizar que sean aplicadas por un examinador calificado y que los resultados se empleen apropiadamente; y (b) impedir una familiaridad general con su contenido, ya que ello invalidaría el instrumento.

Examinador calificado. La necesidad de un examinador calificado se vuelve evidente en cada uno de los tres aspectos principales de la sitùación de prueba: la selección del test, su aplicación y calificación, y la interpretación de los resultados. Los tests no pueden elegirse como se escoge una podadora de un catálogo. No pueden evaluarse por el nombre, el autor u otras señas de identificación. Para estar seguros, no se requiere de entrenamiento psicológico al considerar factores como el costo, el volumen o la facilidad de transportación de los materiales de la prueba, el tiempo requerido para resolverla y la facilidad, así como la rapidez para calificarla. Por lo general, es posible obtener de un catálogo de tests la información sobre esos puntos prácticos que desde luego deben ser considerados al planear un programa de evaluación. Sin embargo, para que una prueba cumpla su función resulta imprescindible evaluar sus méritos técnicos en términos de sus características de validez, confiabilidad, grado de dificultad y normas. Sólo así es posible que los usuarios determinen qué tan adecuado es un instrumento para un propósito en particular o para las personas a las que planean aplicarlo.

Al hablar de la estandarización de las pruebas señalamos ya la importancia de contar con un examinador capacitado. Para que las puntuaciones que obtienen diferentes examinadores sean comparables o para evaluar los resultados de un individuo en tér-

minos de las normas publicadas es necesario percatarse de la importancia de seguir con precisión las instrucciones y de familiarizarse con ellas. También resulta fundamental el control cuidadoso de las condiciones de aplicación. De modo similar, la calificación incorrecta o inexacta puede inutilizar el resultado. Sin los procedimientos adecuados de supervisión, es mucho más probable que ocurran errores de calificación de lo que la gente cree.

La interpretación adecuada de los resultados requiere una comprensión cabal de la prueba, del sujeto que la presentó y de las condiciones en que fue aplicada. Sólo es posible determinar con objetividad lo que se mide si se hace referencia a los procedimientos que la validaron. También es pertinente contar con otra clase de información relativa a la confiabilidad, la naturaleza del grupo con el que se establecieron las normas, etc. Al interpretar los resultados es importante contar con algunos antecedentes del examinado. Distintas personas pueden obtener la misma puntuación por razones muy diferentes, por lo que las conclusiones a las que se llegue también deberían ser disímiles. Por último, también debe prestarse atención a factores especiales que pudieran haber afectado una calificación, como algunas condiciones inusuales de aplicación, el estado emocional o físico del examinado y su grado de experiencia con las pruebas.

La función del usuario. Durante los ochenta y los noventa el reconocimiento de la importante función del usuario constituyó un avance significativo en el campo de las pruebas psicológicas (Anastasi, 1990b). En este contexto, el usuario es cualquiera que utiliza los resultados de una prueba como fuente de información para tomar decisiones prácticas y puede ser, pero no necesariamente, el mismo que la aplica y la califica. Como ejemplos citemos a los maestros, consejeros, administradores de sistemas escolares o de personal en la industria o el gobierno. La mayor parte de las críticas no se dirige a los rasgos intrínsecos a las pruebas, sino al destino que usuarios mal calificados dan a los resultados. El deseo de encontrar atajos, respuestas rápidas y soluciones rutinarias simples para problemas complejos da lugar a algunos abusos. La presión temporal de una sobrecarga de trabajo puede fomentar tales recursos; sin embargo, es probable que la causa más frecuente sea el conocimiento insuficiente o inadecuado de las pruebas (Eyde, Moreland, Robertson, Primoff y Most, 1988; Moreland, Eyde, Robertson, Primoff y Most, 1995; Tyler y Miller, 1986).

En los Estados Unidos, comisiones especiales de organizaciones profesionales trabajan en conjunto con los editores de las pruebas para prevenir su mal uso. Un ejemplo notable es el proyecto del *Test User Qualifications Working Group* ("Grupo de Trabajo para la Certificación de los Usuarios de las Pruebas"), conocido por las siglas TUQWoG (Eyde *et al.*, 1988), cuya meta principal es el desarrollo de una base de datos empíricos de las condiciones esenciales que tienen que cumplir los usuarios de las pruebas y que los editores deben incluir en sus formas de certificación para permitir su adquisición. Luego de una investigación llevada a cabo durante cinco años en ese país, el proyecto TUQWoG formó una impresionante base de datos. Algunos editores ya han empezado a emplear los resultados en las formas de certificación del comprador. Más adelante se formó otro grupo con el propósito de utilizar la base de datos del TUQWoG para elaborar directrices y materiales de capacitación para los usuarios. El primer producto de este nuevo grupo, conocido como TUTWoG, Test User Training Working Group (Grupo de

Trabajo para la Capacitación de los Usuarios de las pruebas), es un libro que hace un recuento de los abusos más comunes con el propósito de prevenirlos (Eyde *et al.*, 1993). Los casos se basan en casos reales de abusos observados en diversos medios y que reveló una encuesta realizada para el proyecto. Moreland *et al.* (1995) presentan un resumen.

Seguridad del contenido de la prueba y comunicación de la información pertinente. Si una persona memorizara las respuestas correctas de una prueba de ceguera al color, la prueba quedaría totalmente invalidada dado que ya no podría ser una medida de su visión cromática. Es evidente que el contenido de las pruebas debe ser restringido para impedir los intentos por falsear los resultados; sin embargo, en otros casos el efecto de la familiaridad puede ser menos evidente o la prueba resultar invalidada de buena fe por personas mal informadas; por ejemplo, un maestro puede hacer que sus alumnos resuelvan problemas muy parecidos a los que presenta un test de inteligencia con el propósito de que "los niños estén preparados para presentarlo". Semejante actitud es simplemente un remanente del procedimiento que suele seguirse al preparar a los alumnos para un examen escolar; pero cuando se aplica a un test de inteligencia es probable que eleve las puntuaciones sin afectar de manera apreciable el área más amplia de conducta que la prueba pretende medir. En tales condiciones, se reduce la validez de la prueba como instrumento predictivo o de diagnóstico.

Garantizar la seguridad del contenido de una prueba no tiene por qué interferir con la comunicación de la información pertinente a las personas evaluadas, los profesionales interesados y el público en general, pues tal comunicación cumple varios objetivos. Primero, tiende a desvanecer cualquier velo de misterio asociado con el uso de las pruebas y en consecuencia ayuda a corregir los conceptos erróneos que prevalecen acerca de su propósito y el significado de sus resultados; para esto, algunos de los editores más importantes distribuyen folletos informativos. El segundo tiene que ver con los procedimientos técnicos seguidos al elaborar y evaluar los instrumentos; esta información ofrece datos importantes acerca de la confiabilidad, la validez y otras propiedades psicométricas del instrumento y, por lo general, se incluye en el manual técnico preparado para cada prueba y está disponible para cualquier persona interesada.

El tercer objetivo de la información consiste en familiarizar a las personas evaluadas con los procedimientos de la prueba, disminuir la ansiedad y lograr que cada una haga su mejor esfuerzo. Para estos fines se han preparado varios folletos explicativos, algunos de naturaleza general y otros para herramientas específicas como el Test de Evaluación Escolar de la Junta Universitaria (*College Board's Scholastic Assessment Test*), materiales que analizaremos en una sección posterior del capítulo. El cuarto objetivo, muy importante, es la retroalimentación que se brinda al examinado sobre su desempeño en la prueba. Los psicólogos han prestado una atención considerable a las formas más útiles y significativas de entregar esa información en diferentes contextos. En los capítulos 17 y 18 examinaremos los procedimientos apropiados.

La difusión de la información acerca de las pruebas es de gran importancia, y existen maneras útiles pero también dañinas de llevar a cabo esta tarea. Un ejemplo se encuentra en los precipitados intentos legislativos realizados en los Estados Unidos por introducir controles gubernamentales tanto a nivel estatal como federal (Bersoff, 1981, 1983; B. Lerner, 1980b). A finales de los setenta, fueron aprobadas leyes estatales que

regulan la divulgación de la información sobre las pruebas en California y en Nueva York. La de este último estado, que es la más extrema, requiere una divulgación estricta de las preguntas y respuestas de las pruebas aplicadas en todos los programas de evaluación a gran escala para la admisión a instituciones de educación superior.

Como semejante requisito supone la preparación de una nueva versión de cada prueba en cada ocasión que se aplica, puede tener varios efectos adversos como la disminución en las fechas de aplicación disponibles durante el año, el aumento en las cuotas que debe pagar el solicitante y la disminución en la calidad de los procedimientos de control al elaborar la prueba y al igualar las puntuaciones de las pruebas aplicadas en diferentes momentos. También es digno de mención que son muy pocas las personas que aprovechan la oportunidad que la legislación sobre divulgación les brinda, y que dicha divulgación no aumenta de manera significativa el desempeño en la segunda aplicación con otra forma de examen (Stricker, 1984). Las metas que impulsaron la promulgación de las leyes se alcanzan de mejor manera, y sin los nocivos efectos colaterales, si se fortalecen los procedimientos disponibles para comunicar la información de las pruebas.

APLICACIÓN DE LA PRUEBA

El fundamento de las pruebas es que pueden generalizar la muestra de conducta observada en la situación de prueba al comportamiento manifestado en otras situaciones. El resultado de una prueba debería ayudarnos a predecir cómo se sentirá y actuará el cliente fuera de la clínica, cuál será el desempeño académico del estudiante y cuál el desempeño laboral de un solicitante. Todas las influencias en la situación de prueba constituyen un error de varianza y reducen su validez. Por eso es tan importante identificar cualquiera que se relacione con la prueba y que pueda limitar o afectar la generalización de sus resultados.

Podríamos dedicar todo un volumen al análisis de los procedimientos deseables de aplicación de las pruebas, pero tal estudio escapa a los propósitos de este libro. Además, resulta más práctico adquirir dichas técnicas en medios específicos ya que, por lo general, no hay quien esté interesado en todas las formas de aplicación, del examen de infantes a las pruebas clínicas de pacientes psicóticos o a la aplicación de programas masivos de evaluación para personal militar. En consecuencia, nuestro análisis se orientará principalmente a los fundamentos de la aplicación de pruebas en lugar de abordar preguntas concretas sobre su puesta en práctica. Un excelente ejemplo de ello puede encontrarse en la concienzuda revisión de la evaluación individual infantil hecha por Sattler (1988, capítulo 5).

Preparativos previos a la aplicación. El requisito más importante de un buen procedimiento de aplicación son los preparativos. Durante la aplicación de la prueba no puede haber emergencias, por lo que tienen que hacerse esfuerzos especiales para anticiparlas e impedirlas. Sólo así se garantiza la uniformidad del procedimiento.

La preparación para la sesión de aplicación adopta muchas formas. En la mayor parte de las pruebas individuales resulta esencial memorizar las instrucciones verbales exactas. Incluso en una prueba de aplicación grupal en la que se leen las instrucciones

a los examinados, familiarizarse con el material previene los errores y dudas durante la lectura y permite un estilo más natural e informal durante la aplicación. Otro paso preliminar importante es la preparación de los materiales, que en las pruebas individuales y especialmente en las de desempeño consiste en disponer todo lo necesario para facilitar su uso con un mínimo de búsqueda o tropiezos. Es conveniente que los materiales se coloquen en un mueble cercano a la mesa en la que se realizará la aplicación de modo que estén al alcance del examinador pero que no distraigan al examinado. Cuando se emplean aparatos complejos puede ser necesario vigilarlos y calibrarlos de manera periódica. En las pruebas de aplicación grupal, todos los cuadernillos, las hojas de respuesta, los lápices especiales, o cualquier otro material, deben ser cuidadosamente contados, verificados y arreglados antes del día de la aplicación de la prueba.

Otro requisito importante, tanto en las pruebas de aplicación individual como en las de grupo, es la familiaridad absoluta con el procedimiento de aplicación. Para las pruebas individuales, suele ser esencial recibir una capacitación supervisada en la aplicación de la prueba. Dependiendo de la naturaleza del instrumento y de las personas examinadas; la capacitación puede requerir desde unas cuantas sesiones de demostración y práctica hasta más de un año de instrucción. Para las pruebas de grupo, en especial en los proyectos a gran escala, la preparación puede incluir instrucciones previas a los examinadores y ayudantes, de modo que todos estén completamente informados sobre las funciones que debe realizar. Por lo general, el examinador lee las instrucciones, se ocupa de llevar el tiempo y está a cargo del grupo en el salón. Los ayudantes entregan y recogen los materiales, se aseguran de que se sigan las instrucciones, responden las preguntas de los examinados dentro de los límites especificados en el manual e impiden las copias.

Condiciones de aplicación. El procedimiento estandarizado se aplica no sólo a las instrucciones verbales, el tiempo y otros aspectos de la prueba, sino también al ambiente. Es necesario elegir un salón adecuado para el examen, el cual debe estar libre de ruidos y distracciones y ofrecer a los examinados condiciones adecuadas de iluminación, ventilación, asientos y espacio de trabajo. También deben tomarse precauciones para prevenir interrupciones durante la aplicación, por lo que es conveniente colocar en la puerta un cartel que indique que se está aplicando la prueba y asegurarse de que todo el personal se haya enterado de que la señal significa que nadie puede entrar bajo ninguna circunstancia. En las pruebas colectivas, puede ser necesario cerrar la puerta o poner a un ayudante afuera para impedir la entrada de los rezagados.

Es importante darse cuenta de las condiciones en que se realiza la prueba ya que éstas pueden influir en sus resultados. Incluso aspectos que parecen menores pueden alterar de manera apreciable el desempeño; por ejemplo, el uso de pupitres fijos o bien de sillas con paleta para el brazo demostró ser significativo en un proyecto de aplicación grupal con estudiantes de secundaria, pues el grupo que utilizó pupitres tendía a obtener mayores calificaciones (T. L: Kelley, 1943; Traxler y Hilkert, 1942). También hay evidencias que demuestran que la hoja de respuesta que se emplee puede influir en los resultados (F. O. Bell, Hoff y Hoyt, 1964). A veces, los examinadores utilizan en sus pruebas colectivas hojas de respuesta diferentes de las que se emplearon en la muestra de estandarización, lo que se debe al establecimiento de oficinas independientes de

calificación de pruebas y procesamiento de datos que entregan sus propias hojas de respuesta, las que pueden ser calificadas por máquinas. Dado que se carece de una verificación empírica, no es posible asumir que las hojas son equivalentes. Al examinar a niños de grados inferiores al quinto, el uso de *cualquier* hoja diferente puede disminuir de forma notable sus calificaciones (Cashen y Ramseyer, 1969; Ramseyer y Cashen, 1971), así que en esos niveles suele ser preferible hacer que marquen las respuestas en el propio cuadernillo de la prueba.

Todavía más significativas a cualquier edad son las diferencias entre la aplicación computarizada o de lápiz y papel de la misma prueba, por ello se ha dedicado considerable atención al efecto que tiene sobre las normas, la confiabilidad y la validez en relación con la naturaleza de la prueba y la población examinada. Lo anterior ha conducido a la formulación de lineamientos profesionales para que los usuarios decidan qué tan comparables son los resultados de las dos aplicaciones (Butcher, 1987; Hofer y Green, 1985).

Se ha demostrado que muchas otras condiciones sutiles afectan el desempeño tanto en los tests de habilidad como en los de personalidad. Que el examinador sea extraño o conocido para el examinado puede suponer una diferencia significativa en los resultados (Sacks, 1952; Tsudzuki, Hata y Kuze, 1957). En otro estudio se comprobó que los modales y la conducta del examinador (sonreír, asentir con la cabeza y hacer comentarios como "bien" o "perfecto"), tienen un efecto decisivo en los resultados (Wickes, 1956). Al aplicar una técnica proyectiva que requiere que el examinado escriba historias que se ajusten a ciertas imágenes, se descubrió que la presencia del examinador en la habitación tendía a inhibir la inclusión de contenido emocional en las historias (Bernstein, 1956). En la aplicación de una prueba de mecanografía, los solicitantes escribían a una tasa significativamente más alta cuando se les probaba solos que cuando el examen se hacía en grupos de dos o más personas (Kirchner, 1966).

Podríamos multiplicar con facilidad los ejemplos, hecho que tiene tres implicaciones. Primero, siga con minucioso detalle los procedimientos estandarizados. Es responsabilidad del autor de la prueba y del editor explicar los procedimientos de manera clara y completa en el manual de la prueba. Segundo, lleve registro de cualquier condición inusual que tenga lugar durante la aplicación, aunque sea menor. Tercero, al interpretar los resultados, tome en consideración las condiciones en las que se llevó a cabo. En la evaluación a fondo de un sujeto, el examinador experimentado ocasionalmente puede apartarse del procedimiento estandarizado para obtener información adicional por razones especiales. En esos casos, ya no se puede interpretar los resultados en términos de las normas de la prueba. En tales circunstancias, los estímulos de la prueba se utilizan únicamente para una exploración cualitativa, y las respuestas tendrían que tratarse de la misma manera que cualquier otra observación informal de la conducta o que los datos obtenidos en una entrevista.

Presentación de la prueba: rapport y orientación del examinado. Al aplicar una prueba, se entiende por *rapport* los esfuerzos del examinador por despertar el interés del examinado, lograr su cooperación y animarlo a responder de manera apropiada a los objetivos del instrumento. En los tests de habilidad, el objetivo requiere una concentración cuidadosa en las tareas presentadas y hacer el mejor esfuerzo por un buen desempe-

ño. En los inventarios autodescriptivos de personalidad, el objetivo es obtener respuestas francas y honestas a las preguntas sobre las conductas cotidianas; en ciertas técnicas proyectivas, se requiere de un informe completo de las asociaciones evocadas por los estímulos sin censura ni manipulación del contenido. Otras pruebas exigen otras aproximaciones, pero en todos los casos, el examinador se esfuerza por motivar al examinado a seguir las instrucciones de manera tan completa y concienzuda como le sea posible.

La capacitación de los examinadores incluye la adquisición de técnicas para el establecimiento de *rapport*, así como de otras que están relacionadas más directamente con la aplicación de la prueba. En el establecimiento del *rapport*, como en otros procedimientos de la prueba, resulta esencial la uniformidad de las condiciones para que los resultados sean comparables. Si una niña recibe un premio cada vez que resuelve un problema de la prueba, su desempeño no puede ser comparado directamente con las normas o con el desempeño de otros niños que sólo son motivados con incentivos o alabanzas verbales. Cualquier desviación de las condiciones motivacionales normales de una prueba tiene que anotarse y considerarse al interpretar la ejecución.

Aunque el *rapport* se establece más plenamente en las pruebas individuales, en las de grupo también es posible emprender acciones para motivar a los examinados y disminuir su ansiedad. Las técnicas varían con la naturaleza de la prueba, la edad y otras características del examinado. Cuando se trabaja con preescolares, deben considerarse factores especiales como la timidez ante los desconocidos, la disposición a distraerse y el negativismo. Un estilo amistoso, cariñoso y relajado de parte del examinador ayuda a darles confianza. El niño tímido y retraído necesita más tiempo para familiarizarse con los alrededores, por lo que es mejor que el examinador no se extienda demasiado al principio, sino que espere a que el niño esté listo para hacer el primer contacto. Los periodos de prueba deben ser breves y las tareas variadas e interesantes para el niño; tienen que presentarse como un juego, y antes de introducir una tarea nueva hay que despertar su curiosidad. A esta edad se requiere cierta flexibilidad de procedimiento por la posibilidad de negativas, pérdida de interés y otras manifestaciones de negativismo.

Los niños en los dos o tres primeros grados de la escuela elemental presentan muchos de los problemas observados en los preescolares, por lo que el método del juego sigue siendo la forma más eficaz de despertar su interés en la prueba. Los mayores pueden ser motivados si se apela al espíritu de competencia y al deseo de hacer un buen trabajo; sin embargo, al evaluar a niños cuyos antecedentes educativos los ponen en desventaja o que provienen de culturas diferentes, el examinador no puede suponer que estarán motivados para destacar en las tareas académicas en el mismo grado que los niños de la muestra de estandarización. En los capítulos 9, 12 y 18 veremos éste y otros problemas relacionados con la evaluación de sujetos con experiencias disímiles.

Es probable encontrar problemas motivacionales especiales al evaluar a individuos con perturbaciones emocionales, prisioneros y delincuentes juveniles, que posiblemente manifiesten actitudes desfavorables como suspicacia, inseguridad, temor o indiferencia cínica, en especial cuando son evaluados en un centro de reclusión. También es probable que ciertas peculiaridades de sus experiencias influyan en su desempeño de manera nociva; por ejemplo, como resultado de los fracasos y las frustraciones sufridos en la escuela pueden haber desarrollado sentimientos de hostilidad e inferioridad hacia las tareas académicas, que se parecen a las pruebas. El examinador

experimentado debe hacer esfuerzos especiales por establecer el *rapport* en tales condiciones. En cualquier caso, debe ser sensible a esas dificultades especiales y tomarlas en consideración al interpretar y explicar el desempeño en la prueba.

Al evaluar a niños escolares o a adultos debe recordarse que cada prueba representa una amenaza implícita para el prestigio del individuo, por lo que es necesario tranquilizarlo desde el inicio; por ejemplo, es útil explicarle que no se espera que nadie termine o responda correctamente todos los reactivos, pues, de otra manera, el examinado experimentaría una sensación de fracaso al avanzar en los reactivos más difíciles o al ver que no es capaz de terminar algún segmento en el tiempo permitido.

Dado que lo inesperado y lo desconocido suelen producir ansiedad, también es deseable eliminar, hasta donde resulte posible, las sorpresas en la situación de prueba. Aunque muchas pruebas colectivas incluyen una explicación preliminar que el examinador lee ante el grupo, un procedimiento aún mejor consiste en proporcionar con antelación a cada examinado materiales que expliquen el propósito y la naturaleza del instrumento, le ofrezcan sugerencias generales sobre cómo presentarla y que contengan algunos reactivos de muestra. Por lo general, quienes participan en programas de evaluación a gran escala tienen acceso a los manuales explicativos, como los llevados a cabo por la Junta Universitaria (*College Board*).

La prueba para adultos implica otros problemas, pues a diferencia de los escolares es poco probable que se esfuercen en una tarea simplemente porque les es asignada. Por ello se hace más importante "venderles" el propósito de la prueba, aunque los estudiantes de educación media y superior también responden a ese llamado. Habitualmente, es posible lograr la cooperación de los examinados al convencerlos de que les conviene obtener una puntuación válida, es decir, un resultado que indique correctamente lo que pueden hacer en lugar de sobrestimar o subestimar sus habilidades. La mayoría de la gente puede entender que una decisión incorrecta, tomada a partir del resultado no válido de la prueba, puede significarle fracasos, pérdida de tiempo y frustraciones. Como los sujetos se dan cuenta de que son ellos los que más tienen que perder, este sistema puede servir no sólo para motivarlos a hacer su mejor esfuerzo en los tests de habilidad, sino también para reducir los engaños y fomentar el reporte honesto en los inventarios de personalidad. Por supuesto, a nadie le conviene ser admitido en un curso para el que no tiene las habilidades o el conocimiento requeridos o ser asignado a un puesto que no puede desempeñar o que encuentra desagradable.

EXAMINADOR Y VARIABLES SITUACIONALES

Periódicamente se publican estudios de gran alcance sobre los efectos que el examinador y las variables situacionales tienen en los resultados de las pruebas (Lutey y Copeland, 1982; Masling, 1960; S. B. Sarason, 1954; Sattler, 1970, 1988; Sattler y Theye, 1967). Aunque se han descubierto algunos efectos en las pruebas objetivas de grupo, la mayor parte de los datos proviene de técnicas proyectivas o de tests de inteligencia. Es más probable que esas variables extrañas operen con estímulos ambiguos y no estructurados, así como con tareas difíciles y novedosas, que con funciones definidas con claridad y bien aprendidas. En general, los niños son más susceptibles que los adultos a los efectos de las variables situacionales y del aplicador, por lo que la función

de éste resulta especialmente importante al evaluar a preescolares. La probabilidad de que estas condiciones tengan algún influjo también es mayor en las personas inseguras o con trastornos emocionales de cualquier edad.

Se ha investigado la relación entre el desempeño en tests de inteligencia y técnicas proyectivas aplicadas individualmente con muchas variables del examinador, entre las que se incluyen edad, sexo, raza, posición profesional o socioeconómica, capacitación y experiencia, características de personalidad y apariencia. Aunque se han encontrado varias relaciones significativas, los resultados suelen ser erróneos o no concluyentes porque el diseño experimental no logró controlar o aislar la influencia de diferentes características del aplicador o del examinado, lo que supone la posible confusión de los efectos de dos o más variables.

Lo que se ha demostrado con mayor claridad es que la conducta del aplicador antes y durante la prueba puede alterar los resultados; por ejemplo, en investigaciones controladas se ha descubierto diferencias notables en el desempeño en un test de inteligencia como resultado de una relación interpersonal "cálida" frente a una "fría" entre aplicador y examinado, o un estilo del aplicador rígido y reservado frente a uno más natural (Exner, 1966; Masling, 1959). Más aún, puede haber interacciones significativas entre las singularidades del aplicador y examinado en el sentido de que las mismas características del aplicador, o su estilo de aplicar la prueba, pueden tener un efecto muy diferente en distintos examinados en función de la propia personalidad del examinado. Interacciones similares pueden ocurrir con las variables de la tarea, como la naturaleza de la prueba, el propósito de su aplicación y las instrucciones dadas a los evaluados. Dyer (1973) ha agregado otras variables a esta lista, y llama la atención sobre la posible influencia de las diversas percepciones que aplicadores y examinados tienen de las funciones y objetivos de la prueba.

Las expectativas del aplicador constituyen otra forma en que éste puede afectar sin quererlo las respuestas del examinado. Se trata simplemente de un caso especial de la profecía autorrealizada (Harris y Rosenthal, 1985; R. Rosenthal, 1966; R Rosenthal y Rosnow, 1969). Un ejemplo se encuentra en un experimento realizado con el test de Rorschach (Masling, 1965). Los aplicadores fueron 14 estudiantes graduados que se ofrecieron como voluntarios; a siete de ellos se les dijo, entre otras cosas, que los aplicadores experimentados provocaban más respuestas humanas que animales, mientras que a los otros siete se les informó lo contrario. En tales condiciones, los dos grupos obtuvieron de sus examinados proporciones significativamente diferentes de respuestas animales o humanas, y esas diferencias ocurrieron a pesar de que ni los aplicadores ni los examinados dijeron estar conscientes de ninguna influencia. Más aún, las grabaciones de las sesiones no revelaron evidencias de influencias verbales por parte de ningún aplicador. Según parece las expectativas de los aplicadores operaron con sutiles claves posturales y faciales a las que respondieron los examinados.

Además del aplicador, otros aspectos de la situación de prueba pueden influir significativamente en el desempeño; por ejemplo, a menudo se evalúa a los reclutas al poco tiempo de su inducción, durante un periodo de intenso reajuste a una situación desconocida y estresante. En una investigación destinada a probar el efecto de aclimatarse a dicha situación sobre el desempeño en la prueba, se aplicó la Batería de Clasificación de la Marina (*Navy Classification Battery*) a 2 724 reclutas en su noveno día

en el Centro de Entrenamiento Naval (L. V. Gordon y Alf, 1960). Al comparar sus resultados con los de 2 180 reclutas probados en el momento habitual, al tercer día en el centro, el primer grupo obtuvo calificaciones superiores en todos los subtests de la batería.

Las actividades que realizan los sujetos justo antes de la prueba también tienen un efecto en su desempeño, en particular si producen perturbación emocional, fatiga u otras condiciones desventajosas. En una investigación realizada con niños de tercer y cuarto grado se hallaron evidencias de que la actividad que había tenido lugar antes en el aula influía en el CI que obtenían en la Prueba de Dibujo de un Hombre (McCarthy, 1944). En una ocasión, los estudiantes habían estado ocupados en la redacción de una composición sobre "Lo mejor que me ha sucedido", mientras que en la segunda escribían sobre "Lo peor que me ha pasado". El CI de la segunda prueba, después de lo que puede haber sido una experiencia depresiva, era en promedio cuatro o cinco puntos inferior al de la primera prueba. Estos descubrimientos fueron corroborados en otra investigación diseñada en concreto para determinar el efecto de la experiencia inmediatamente precedente sobre la Prueba de Dibujo de un Hombre (*Draw-a-Man Test*) (Reichenberg-Hackett, 1953). En este estudio, los niños que habían pasado por la experiencia gratificante de resolver un acertijo interesante y recibir por recompensa dulces y juguetes, mostraron mejores puntuaciones que quienes habían tenido una experiencia neutral o menos agradable. W. E. Davis (1969a, 1969b) obtuvo resultados similares con estudiantes universitarios. El desempeño en una prueba de razonamiento aritmético fue significativamente inferior cuando la precedía una experiencia fallida en una prueba de comprensión verbal que en un grupo de control al que no se aplicó dicha prueba o en otro que había presentado una prueba estándar de comprensión verbal en condiciones ordinarias.

Diversos estudios se han interesado por los efectos de la retroalimentación en los resultados de la prueba en la ejecución subsecuente del individuo. En una investigación muy bien diseñada con estudiantes de séptimo grado, Bridgeman (1974) encontró que el desempeño en una prueba subsecuente era bastante mejor cuando se recibía retroalimentación de "éxito" que cuando era de "fracaso" en una prueba inicial en la que la ejecución de ambos grupos había sido igualmente buena. La retroalimentación motivacional funciona sobre todo mediante las metas que los participantes se imponen para su desempeño posterior, por lo que es otro ejemplo de la profecía que se autorrealiza; sin embargo, no debe confundirse esta retroalimentación motivacional con la correctiva, con la que se informa al individuo de los reactivos específicos que contestó mal y recibe educación de regularización. En estas condiciones, es mucho más probable que la retroalimentación mejore la ejecución de quienes habían obtenido un mal resultado.

Los ejemplos citados ilustran la gran diversidad de variables relacionadas con las pruebas que pueden alterar los resultados. En la mayor parte de los programas de evaluación bien aplicados, la influencia de esas variables es insignificante para todo propósito práctico; no obstante, el examinador experimentado siempre está en guardia para detectar la operación de dichas variables y disminuir su influjo. Cuando las circunstancias no permiten el control de estas condiciones, resulta necesario restringir las conclusiones derivadas del desempeño en la prueba.

EL PUNTO DE VISTA DEL EXAMINADO

Ansiedad ante la prueba. Entre las primeras investigaciones sobre las reacciones de los evaluados ante la situación de prueba se encuentran las que estudian la ansiedad que produce la prueba. Es indudable que su notoriedad y sus efectos nocivos sobre el desempeño estimularon el interés por esta información. En la aplicación de las pruebas, muchas de las prácticas destinadas a lograr el *rapport* también reducen la ansiedad, lo mismo que los procedimientos que tienden a disminuir las sorpresas y la extrañeza de la situación de prueba y a tranquilizar y animar al examinado. El estilo del aplicador y una sesión bien organizada contribuyen al mismo fin.

Las diferencias individuales en cuanto a la ansiedad que causan las pruebas han sido estudiadas lo mismo en escolares que en universitarios (Gaudry y Spielberger, 1974; Hagtvet y Johnsen, 1992; I. G. Sarason, 1980; Spielberger, 1972). Buena parte de esta investigación fue iniciada por S. B. Sarason y sus colaboradores en Yale (Sarason, Davidson, Lighthall, Waite y Ruebush, 1960). El primer paso consistió en formular un cuestionario para evaluar las actitudes del individuo hacia la prueba; por ejemplo, la forma para los niños contenía reactivos como los siguientes:

¿Te preocupas mucho antes de presentar una prueba?

¿Sientes que tu corazón empieza a latir más aprisa cuando la maestra dice que va a averiguar qué tanto has aprendido?

Cuando estás presentando una prueba ¿piensas que no lo estás haciendo bien?

El principal interés es el descubrimiento de que tanto los resultados de las pruebas de rendimiento académico como los de los tests de inteligencia arrojaron correlaciones negativas con la ansiedad producida por las pruebas. En los estudiantes universitarios se han encontrado correlaciones similares (I. G. Sarason, 1961). Del mismo modo, estudios longitudinales revelaron una relación inversa entre los cambios en el grado de ansiedad y los cambios en el desempeño en tests de inteligencia o de aprovechamiento (K. T. Hill y S. B. Sarason, 1966; S. B. Sarason, K. T. Hill y Zimbardo, 1964).

Por supuesto, estos descubrimientos no indican la presencia de relaciones causales. Es posible que los estudiantes se sientan ansiosos con las pruebas porque suelen tener un mal desempeño al realizarlas y por ende han pasado por frustraciones y fracasos. En apoyo a esta interpretación está el hecho de que en los subgrupos de mayores puntuaciones en los tests de inteligencia desaparece la correlación negativa entre el grado de ansiedad y el desempeño (Denny, 1966; Feldhusen y Klausmeier, 1962). Por otro lado, se cuenta con evidencias de que al menos parte de esta relación proviene de los efectos nocivos de la ansiedad sobre el desempeño en las pruebas. En un estudio (Waite, Sarason, Lighthall y Davidson, 1958), niños con diferentes grados de ansiedad (elevada y baja), pero con resultados similares en los tests de inteligencia, hicieron varios ensayos en una tarea de aprendizaje. Aunque al principio su ejecución en la prueba de aprendizaje era igual, el grupo con menor ansiedad mostró una mejora significativamente mayor que el grupo ansioso.

Varios investigadores han comparado la ejecución en las pruebas en condiciones diseñadas para producir un estado "ansioso" o uno "relajado"; por ejemplo, Mandler y

Sarason (1952) descubrieron que instrucciones que se referían al ego (como decir a los examinados que se esperaba que todos terminaran en el tiempo permitido) tenían un efecto benéfico sobre la ejecución de los más tranquilos, pero uno nocivo en los ansiosos. Otros estudios también han encontrado una relación entre las condiciones de la prueba y las características individuales como el nivel de ansiedad y la motivación para el logro (Lawrence, 1962, Paul y Eriksen, 1964). Parece probable que la relación entre ansiedad y ejecución no sea lineal, es decir, que una poca de ansiedad sea benéfica y que mucha sea nociva. Los individuos que habitualmente son poco ansiosos se benefician de las condiciones de la prueba que generan cierta ansiedad, mientras que los que suelen ser presa del nerviosismo trabajan mejor en condiciones más relajadas.

No cabe duda de que una ansiedad elevada crónica ejerce un efecto nocivo sobre el aprendizaje académico y el desarrollo intelectual. La ansiedad interfiere con la adquisición y la recuperación de la información (Hagtvet y Johnsen, 1992). Sin embargo, es necesario distinguir este efecto del impacto que la ansiedad tiene sobre la prueba (es decir, la medida en que la ansiedad que produce hace que el desempeño del individuo sea poco representativo de su desempeño en otras situaciones), que es lo que constituye el objeto de nuestro estudio. Debido a la presión competitiva que experimentan los estudiantes cuyo ingreso a la universidad está próximo, se ha argumentado que el desempeño en las pruebas de admisión puede resultar muy afectado por la ansiedad que causan. En una investigación cuidadosa y bien diseñada, French (1962) comparó el desempeño de estudiantes de bachillerato en un examen aplicado como parte de la aplicación regular del Test de Aptitud Académica (*Scholastic Aptitude Test*) con su ejecución en una forma paralela de esa prueba aplicada en otro momento en condiciones "relajadas". Las instrucciones en este último caso especificaban que la prueba se aplicaba con propósitos de investigación y que los resultados no se enviarían a ninguna universidad. Ahora bien, éstos mostraron que la ejecución no fue peor durante la aplicación estándar que durante la relajada. Más aún, la validación de los puntajes de la prueba con las notas escolares no difería en las condiciones. Varias investigaciones recientes también han cuestionado el estereotipo común del estudiante ansioso por la prueba que sabe la materia pero que se "paraliza" al presentar el examen (véase Culler y Holahan, 1980). En esta investigación se descubrió que los estudiantes que calificaban más alto en una escala de ansiedad obtenían en promedio menores calificaciones y solían tener peores hábitos de estudio que los que habían calificado bajo.

La investigación sobre la naturaleza, la medición y el tratamiento de la ansiedad producida por los exámenes ha continuado a paso creciente (I. G. Sarason, 1980; Spielberger, Anton y Bedell, 1976; Spielberger, González y Fletcher, 1979; Spielberger, González, Taylor, Algaze y Anton, 1978; G. S. Tryon, 1980) y ha permitido identificar dos componentes importantes relacionados con la naturaleza de la ansiedad, la emocionalidad y la preocupación. El componente de emocionalidad comprende sentimientos y reacciones fisiológicas, como tensión y aumento del ritmo cardiaco. La preocupación, o componente cognoscitivo, incluye pensamientos negativos sobre sí mismo, como la expectativa de fracaso y el temor por sus consecuencias. Esos pensamientos desvían la atención de la conducta orientada a la tarea que la prueba demanda y, por consecuencia, trastornan el desempeño. Diversos inventarios de ansiedad miden ambos componentes, y aunque son de uso común en la investigación, hasta hace muy poco

sólo estaban disponibles en los informes de la bibliografía empírica. El Inventario de Ansiedad (*Test Anxiety Inventory*) elaborado por Spielberger y sus colaboradores es un ejemplo que explicamos en el capítulo 13 y que aparece en el apéndice A.

Se ha dedicado un considerable esfuerzo al desarrollo y la evaluación de métodos para el tratamiento de la ansiedad producida por las pruebas que incluyen varios procedimientos de terapia conductual (capítulo 17) para reducir el componente emocional. Los resultados han sido en general buenos, pero debido a las fallas metodológicas de los estudios de evaluación resulta difícil atribuir la mejoría a alguna técnica (G. S. Tryon, 1980). De hecho, este componente en el test de ansiedad tiende a disminuir del test al retest incluso en los grupos de control sin intervención terapéutica, así como en grupos de control especiales que recibieron una pseudoterapia creíble. Más aún, su reducción tuvo poco o ningún efecto sobre el nivel de desempeño.

Cuando el tratamiento se dirige a las reacciones cognoscitivas personales es más probable que mejore el desempeño tanto en las pruebas como en el trabajo escolar. La investigación disponible hasta ahora indica que los mejores resultados se obtienen al combinar programas de tratamiento para eliminar la emocionalidad y la preocupación así como la mejora de los hábitos de estudio. La ansiedad producida por los exámenes es un fenómeno complejo con causas múltiples, y la contribución relativa de cada causa varía con el individuo. Para que funcionen, los programas de tratamiento tendrían que adaptarse a las necesidades individuales. También debe reconocerse que esta ansiedad es sólo una manifestación de un conjunto más general de condiciones que reducen la eficacia del individuo para aprender.

Investigación amplia de las opiniones de los examinados. Aunque la ansiedad producida por los exámenes es una parte evidente e importante de la conducta de quien presenta una prueba, hay muchos otros elementos que pueden ser estudiados con provecho. Un libro editado en 1993 por Baruch Nevo y R. S. Jäger representa un esfuerzo notable por reunir la información disponible sobre las reacciones de los examinados a las pruebas en medios educativos, industriales, clínicos y de orientación. En los 15 capítulos redactados por investigadores reconocidos sobre diversos aspectos y aplicaciones de las pruebas se encuentran lo mismo informes de publicaciones internacionales sobre cada tema como los descubrimientos de los propios autores. El resultado es un intento serio y fundado por resolver cuestiones que hasta ahora han sido tratadas sobre todo en entornos periodísticos, políticos o legales. El libro funge así como correctivo para las opiniones posiblemente sesgadas y conflictivas sobre las pruebas que abundan en la actualidad; por ejemplo, el primer capítulo incluye 10 excelentes encuestas de opinión sobre las actitudes hacia las pruebas de una amplia gama de poblaciones. Los resultados revelan discrepancias entre las opiniones del público y las afirmaciones de algunos voceros muy publicitados pero poco representativos.

Los capítulos cubren numerosos temas. Algunos explican el desarrollo y el uso de cuestionarios de retroalimentación y las entrevistas de grupo para evaluar las actitudes de diferentes grupos de examinados hacia determinada prueba y sus percepciones sobre lo que ésta mide. En un capítulo comparó las opiniones de estudiantes sobre los exámenes escolares de ensayo y de opción múltiple, que fueron los favorecidos por los resultados. Algunos autores estudian las reacciones de los solicitantes de empleo hacia la

equidad de las pruebas y su relación con el trabajo. Como resultado de sus descubrimientos, varios capítulos sugieren formas de mejorar la aplicación y el ambiente de las pruebas. En conjunto, los capítulos abren a la exploración un área prometedora para buscar soluciones a algunos de los problemas sociales y prácticos de las pruebas, además de que brindan los medios para aumentar la comprensión recíproca de los usuarios de las pruebas y los examinados.

EFECTOS DEL ENTRENAMIENTO SOBRE EL DESEMPEÑO EN LA PRUEBA

Al evaluar los efectos del entrenamiento o la práctica sobre los resultados de las pruebas, una pregunta fundamental es si la mejora se limita a los reactivos que incluye o si se extienden al área más amplia de conducta que la prueba pretende evaluar (Anastasi, 1981a, 1981b). La respuesta a esta pregunta muestra la diferencia entre preparación y educación. Obviamente, cualquier experiencia educativa, formal o informal, dentro o fuera de la escuela, debería reflejarse en el desempeño en las pruebas que estudian los aspectos pertinentes del comportamiento. Dichas influencias no invalidan la prueba en la medida en que su resultado presenta una imagen exacta de la posición del individuo en las habilidades consideradas. Por supuesto, la diferencia es de grado. Las influencias no pueden clasificarse como restringidas o amplias sino que varían en su alcance, de las que afectan una única aplicación de una sola prueba, a las que alteran el desempeño en todos los reactivos de cierta clase y las que influyen en el desempeño del individuo en casi todas las actividades; sin embargo, desde el punto de vista de un examen eficaz, es posible hacer una distinción útil. Así, puede afirmarse que el resultado de la prueba sólo es inválido cuando una experiencia particular eleva la puntuación sin modificar en forma apreciable el área de conducta que pretende medir.

Preparación. Los efectos de la preparación sobre los resultados de las pruebas han sido objeto de una amplia investigación. Los psicólogos británicos realizaron varios estudios que se refieren en especial a los efectos de la práctica y la preparación sobre las pruebas que solían utilizarse para asignar a los niños de 11 años a diferentes escuelas secundarias (Yates *et al.*, 1953-1954). Como era de esperarse, el grado de mejoramiento dependía de la habilidad del examinado, de sus experiencias educativas, de la naturaleza de las pruebas y de la cantidad y clase de preparación recibida. Los individuos con antecedentes educativos deficientes tenían mayores probabilidades de beneficiarse de la preparación especial que quienes habían tenido mejor educación y ya estaban preparados para desempeñar un buen papel en las pruebas. También es evidente que entre más estrecho fuera el parecido entre el contenido de la prueba y el material empleado en la preparación, mayor sería la mejora en los resultados. Por otro lado, entre más se restrinja la instrucción al contenido de la prueba, menos probable es que la mejora se extienda al desempeño de criterio. Más aún, muchos estudios sobre la preparación han arrojado resultados ambiguos y difíciles de interpretar debido a serias deficiencias metodológicas (Anastasi, 1981a; Bond, 1989; Messick, 1980a), entre las cuales sobresale la incapacidad para emplear un grupo de control sin preparación que sea verdaderamente equiparable al grupo preparado; por ejemplo, los estudiantes que se inscriben en los programas comer-

ciales de preparación son autoseleccionados y su habilidad inicial, motivación y otras características personales que influyen en el desempeño de la prueba tienden a diferir de los del grupo de control. Además, en los diseños experimentales que emplean pretest y postest es difícil asegurar que los examinados estén igualmente motivados para hacer un buen trabajo en ambas ocasiones, lo que es cierto sobre todo si una prueba tiene una aplicación regular y la otra una especial con propósitos de práctica o de investigación.

La Junta Universitaria de Exámenes de Admisión a la Universidad (*College Entrance Examination Board*) se encuentra preocupada por la proliferación de cursos comerciales que ofrecen preparar a los estudiantes que aspiran a ingresar en las universidades. Para aclarar el tema, ha realizado diversos experimentos bien controlados para determinar el efecto de la preparación sobre el Test de Aptitud Académica (*Scholastic Aptitude Test*, SAT) y ha revisado los resultados de estudios similares realizados por investigadores independientes (Donlon, 1984; Messick, 1980a, 1981; Messick y Jungeblut, 1981). Los estudios cubren numerosos métodos de preparación e incluyen a estudiantes de escuelas públicas y privadas, así como de grupos minoritarios de áreas urbanas y rurales. La conclusión general es que es poco probable que el ejercicio intenso en reactivos similares a los del SAT produzca ganancias mayores que las que se observan cuando éste se aplica nuevamente luego de un año de instrucción regular.

También debemos decir que en sus procedimientos de elaboración de instrumentos, organizaciones como la Junta Universitaria (*College Board*) y el Consejo de Exámenes de Registro de Graduados (*Graduate Record Examination Board*) investigan la susceptibilidad a la preparación de nuevos reactivos (Evans y Pike, 1973; Powers, 1983; Powers y Swinton, 1984; Swinton y Powers, 1985). En las formas operacionales de la prueba no se conservan los reactivos cuya ejecución puede mejorar mucho con el ejercicio o instrucción a corto plazo y que tienen una naturaleza sumamente restringida. Un ejemplo evidente es el problema que requiere una solución sencilla y perspicaz que, una vez alcanzada, puede aplicarse tal cual a la solución de problemas similares. Cuando vuelven a aparecer, los problemas recuerdan la prueba más que las habilidades de solución de problemas. Otro ejemplo se encuentra en los reactivos complejos que utilizan material novedoso o desconocido y requieren instrucciones largas y complicadas (Powers, 1986).

En el sentido tradicional, la preparación pretende desarrollar habilidades muy restringidas que pueden ser de poca utilidad en las actividades cotidianas. De modo similar, la práctica de "enseñar la prueba" tiende a concentrarse en la muestra particular de habilidades y conocimientos que cubre más que en el área general de conocimientos que la prueba pretende evaluar. Las llamadas leyes de divulgación u "honestidad de las pruebas" que requieren la publicación general de las formas utilizadas luego de una sola aplicación también favorecen la concentración en las habilidades específicas de la prueba, cuya aplicabilidad es limitada. Por último, en la medida en que la preparación sólo está al alcance de algunos, tiende a introducir diferencias individuales en las habilidades específicas de presentación de exámenes, lo que reduce la capacidad de diagnóstico del instrumento.

Perfeccionamiento en las pruebas. A este respecto, también son notables los efectos del perfeccionamiento en las pruebas, o la mera práctica de presentarlas. En estudios con versiones de la misma prueba se observa una tendencia a que la segunda calificación sea más alta. Se ha informado de ganancias significativas en promedio al administrar formas alternas en sucesión inmediata o después de lapsos que van de un

día a tres años (Donlon, 1984; Droege, 1966; Peel, 1951, 1952). Se han obtenido resultados similares con niños normales y sobredotados, estudiantes de educación media y superior y con muestras de empleados. El manual de la prueba debe ofrecer datos sobre la distribución de ganancias esperadas en el retest con una forma paralela, y hay que descontarlas al interpretar los resultados.

Las ganancias no se limitan a las formas alternas, los que tienen mucha experiencia en la presentación de pruebas estandarizadas disfrutan de cierta ventaja sobre quienes presentan la prueba por primera vez (Millman, Bishop y Ebel, 1965; Rodger, 1936). Parte de esta ventaja se debe a que han superado la sensación inicial de extrañeza y a que han adquirido más confianza y mejores actitudes hacia las pruebas, pero también es resultado de cierto traslape en los contenidos y las funciones de muchas pruebas. La familiaridad con algunos reactivos comunes y la práctica en el uso de las hojas de respuesta objetiva también pueden mejorar ligeramente el desempeño. Al comparar los resultados de sujetos con distintos grados de experiencia en las pruebas es importante tomar en cuenta este perfeccionamiento. Para las pruebas aplicadas por computadora debe prestarse atención a la familiaridad del examinado con esta forma de administración (Hofer y Green, 1985).

Las condiciones de perfeccionamiento pueden ser equiparadas de manera eficaz por medio de una breve orientación y sesiones de práctica (Anastasi, 1981a; Wahlstrom y Boersman, 1968). Esta familiarización reduce los efectos de las diferencias previas en la experiencia con las pruebas. La disminución de las diferencias, que son particulares de la situación de prueba, debería permitir una evaluación más válida del área general de conducta que la prueba pretende medir. Tal planteamiento lo ilustra la publicación de la Junta Universitaria titulada *Taking the SAT I: Reasoning Test* ("La presentación del SAT I: Prueba de razonamiento"), un cuadernillo distribuido a todos los aspirantes a la educación universitaria que se registran para presentar esta prueba y que ofrece consejos para prepararla en forma eficaz, ilustra y explica los diferentes reactivos que incluye y reproduce una forma completa de la misma, con la sugerencia a los estudiantes de resolverla en las condiciones normales de tiempo y de calificarla con la clave que se les proporciona. Un cuadernillo similar, *Taking the SAT II: Subject Tests* ("La presentación del SAT II: Pruebas temáticas"), ilustra y explica los reactivos de diferentes pruebas temáticas.

Los Exámenes de Registro de Graduados (*Graduate Record Examinations*, GRE) también proporcionan materiales para familiarizarse con las pruebas. El *Information Bulletin* distribuido a todos los solicitantes comprende la explicación de una muestra de reactivos de la Prueba General (*General Test*), así como una prueba completa previamente aplicada con su clave de calificación. Para presentarla, se publican en un libro (*Practicing to Take the GRE General Test*) formas adicionales y también se dispone de cuadernillos prácticos similares que presentan pruebas individuales del GRE sobre diversas áreas temáticas.

El resurgimiento de los materiales de familiarización aparecidos en los ochenta y los noventa no se limita a los medios impresos, sino que incluye transparencias, diapositivas, películas, videocasetes y *software* para computadora. El Servicio de Pruebas Educativas (*Educational Testing Service*) ha realizado y distribuido muchos de estos materiales, y diseñó algunos para usarse con pruebas específicas, como es el caso de las diapositivas que acompañan a *Taking the SAT* y otros sobre la interpretación de los resultados del SAT y sobre las pruebas de logros de la Junta Universitaria. También se dispone de un programa de computadora para ayudar a comprender las puntuaciones

del SAT, y se ha elaborado un paquete de *software* para los estudiantes que quieren presentar la Prueba General del GRE. Por medio de un programa interactivo, el paquete contiene reactivos de muestra, una situación simulada de supervisión del tiempo, explicaciones de las preguntas contestadas de manera incorrecta y un análisis de los puntos fuertes y débiles del examinado.

Otros materiales (impresos, paquetes de multimedios, *software* para computadora) fueron diseñados para una orientación más general, y cubren temas que van de niños de escuela primaria a adultos. Un ejemplo es *On Your Own: Preparing for a Standardized Test* (1987), un videodisco para uso individual o grupal de estudiantes de secundaria. Una guía sencilla y completa en forma de libro es *How to Take a Test: Doing Your Best* (Dobbin, 1984). También las editoriales comerciales y algunas dependencias gubernamentales de los Estados Unidos han preparado guías para presentar las pruebas, como, por ejemplo, el conjunto de materiales para ser usados con la Batería de Pruebas de Aptitudes Generales (*General Aptitude Test Battery*, GATB) publicado por el Servicio de Empleo de los Estados Unidos.

Instrucción en habilidades cognoscitivas generales. Algunos investigadores han explorado el planteamiento opuesto a la mejora en el desempeño en la prueba. Su meta es la adquisición de habilidades intelectuales de gran aplicación, hábitos de trabajo y estrategias para la resolución de problemas. Los efectos de tales intervenciones deberían manifestarse *lo mismo* en los resultados de las pruebas que en el desempeño de criterio, como los cursos universitarios. De acuerdo con la distinción que presentamos al inicio de la sección, este programa está destinado a brindar educación más que preparación. Algunos de los investigadores que se ocupan del campo han estado trabajando con niños y adolescentes retardados educables (Babad y Budoff, 1974; Belmont y Butterfield, 1977; A. L. Brown, 1974; Budoff y Corman, 1974; Campione y Brown, 1979, 1987; Feuerstein, 1979, 1980; Feuerstein, Rand, Jensen, Kaniel y Tzuriel, 1987), mientras que otros han concentrado sus esfuerzos en estudiantes de educación media y superior con antecedentes de desventajas educativas (Linden y Whimbey, 1990; Whimbey, 1975, 1977, 1980).

Muchos de los procedimientos de capacitación empleados en esos programas fueron diseñados para desarrollar una conducta eficaz de solución de problemas: el análisis cuidadoso de los problemas o bien las preguntas, la consideración de todas las alternativas, los detalles pertinentes y las implicaciones de llegar a una solución, la formulación o elección de una solución deliberada más que impulsiva y la aplicación de criterios elevados para evaluar el propio desempeño. Se trata de estrategias que obviamente deberían mejorar el funcionamiento intelectual no sólo en las pruebas, sino también en el trabajo académico y en muchas otras actividades cotidianas que dependen del aprendizaje escolar; sin embargo, hay una pregunta crucial que tiene que ver con el grado de transferencia y generalización de los efectos aparte de los contenidos y los medios utilizados en la capacitación. Los resultados hasta ahora son prometedores, pero los programas aún se encuentran en etapa de exploración y se requiere de más investigaciones para establecer la amplitud y durabilidad de las mejoras alcanzadas.

Recapitulación. Hemos considerado tres formas de capacitación para las pruebas cuyos objetivos difieren considerablemente. ¿Cómo influyen en la validez de las prue-

bas y en su utilidad práctica como instrumentos de evaluación? La primera forma de entrenamiento es la preparación, en el sentido de un ejercicio intenso y masivo con reactivos similares a los de la prueba. Vimos que las pruebas bien elaboradas eligen aquellos que sean menos susceptibles a dicho ejercicio y los protegen. En tanto que semejante preparación pueda mejorar el desempeño en la prueba, lo hará sin mejorar la conducta de criterio, por lo que la validez del instrumento se reduce y se convierte en una medida menos eficaz de las habilidades generales que pretende evaluar y una forma menos precisa de indagar si el individuo ha adquirido las habilidades y los conocimientos que se requieren para tener éxito en la situación de criterio.

Por otro lado, los procedimientos de orientación sobre la prueba están diseñados para descartar o igualar las diferencias en las experiencias previas a su presentación. Del mismo modo que los efectos de la preparación, estas diferencias representan condiciones que influyen en los resultados de la prueba sin reflejarse necesariamente en el área general de conducta que pretende evaluarse, de ahí que los procedimientos de orientación aumenten la validez de los instrumentos al reducir la influencia de los factores relacionados con las pruebas.

Por último, la preparación en habilidades cognoscitivas de gran aplicación, cuando es eficaz, debe mejorar la habilidad del individuo para enfrentar las tareas intelectuales. Este progreso puede y debe reflejarse en los resultados de la prueba. En la medida en que mejoren tanto los resultados de la prueba como el desempeño en el criterio, esta preparación no modificará la validez de las pruebas y sí aumentará las posibilidades del individuo de alcanzar las metas deseadas.

FUENTES DE INFORMACIÓN

Las pruebas psicológicas se encuentran en un estado de rápido cambio. Las orientaciones se desplazan, hay una corriente constante de nuevas pruebas, de formas revisadas de pruebas antiguas y datos nuevos que pueden refinar o alterar la interpretación de los resultados. El ritmo acelerado de cambio, aunado al vasto número de pruebas existentes, hace que resulte poco práctico tratar de revisar pruebas concretas en un solo texto. En los libros que tratan del uso de los instrumentos en campos como la consejería, la práctica clínica, la selección de personal y la educación es posible encontrar una cobertura más amplia de los instrumentos y los problemas que enfrentan áreas especiales. En los capítulos respectivos anotamos las referencias a estas publicaciones.

De cualquier forma, todo el que trabaje con instrumentos de medición psicológicos necesita familiarizarse con las fuentes de información más directas para mantenerse al corriente. Una de las más importantes es el *Mental Measurements Yearbook* ("Anuario de medición mental") o MMY, establecido y editado por Oscar K. Buros desde 1938 (a partir de 1985, lo publica el Instituto Buros de Medición Mental de la Universidad de Nebraska). Esta serie de anuarios cubre casi todas las pruebas psicológicas, educativas y vocacionales publicadas en inglés y que están comercialmente disponibles. La cobertura es en especial completa para las pruebas de lápiz y papel. Cada anuario incluye las pruebas publicadas durante cierto periodo, por lo que no sustituye a los anuarios anteriores, sino que los complementa. Las primeras publicaciones de esta serie eran simples bibliografías de tests, pero desde 1938 adoptó su forma

actual, que incluye reseñas críticas de las pruebas a cargo de uno o más expertos, así como una lista completa de las referencias publicadas correspondientes a cada prueba. También proporciona información general sobre la editorial, el precio, las formas y la edad de los sujetos para los que es adecuada. El plan actual consiste en publicar un nuevo MMY cada dos o tres años y un suplemento entre cada anuario.

Las entradas del MMY, junto con las reseñas críticas, ahora están disponibles electrónicamente por medio de *SilverPlatter* (véase el apéndice B). La base de datos comienza con las entradas al noveno MMY y se actualiza cada seis meses. Otra publicación del Instituto Buros es *Tests in Print*, ahora en su cuarto volumen (TIP-IV, 1994), editado por L. L. Murphy, Conoley e Impara. Esta publicación proporciona una cobertura acumulada de todas las pruebas que se publican en inglés, junto con la información real y listas de referencias. Cada edición sucesiva también sirve como índice de todos los MMY anteriores.

Otra fuente importante de información sobre las pruebas publicadas son las *Test Collection Bibliographies*, preparadas por el Servicio de Pruebas Educativas (*Educational Testing Service*, ETS), que ofrece una bibliografía actualizada de las pruebas disponibles en áreas específicas de contenido. La cobertura es amplia y comprende todo tipo de pruebas, así como los instrumentos diseñados para usos particulares y poblaciones especiales, como los que padecen discapacidades físicas. Cada entrada contiene información real que incluye datos sobre el autor, la fecha de publicación, la editorial, la población objetivo, el propósito de la prueba y cualquier subpuntaje o variable por medir. Es posible adquirir las bibliografías de pruebas para áreas particulares por un costo nominal en *Test Collection*, ETS (la dirección aparece en el apéndice B). Ésta es una de varias publicaciones del ETS que brinda información actualizada sobre las pruebas y su aplicación.

Además de las pruebas publicadas, también hay una enorme cantidad de ellas descritas o reproducidas en libros, publicaciones periódicas o informes no publicados. De interés especial para los investigadores, estas pruebas han sido examinadas en diversos compendios (por ejemplo, Goldman y Mitchell, 1995). La información actualizada sobre las pruebas no publicadas se encuentra en *Tests in Microfiche*, distribuida por Test Collections, ETS. Cada año se agrega un nuevo conjunto de pruebas y se puede conseguir un índice de cada uno. Los usuarios calificados pueden adquirir pruebas aisladas o conjuntos. El directorio científico de la Asociación Psicológica Estadounidense (*Finding Information*, 1995), una fuente que se actualiza regularmente, tiene una guía clara y concisa para encontrar información sobre pruebas publicadas y no publicadas, y cualquiera que solicite una copia recibe automáticamente la versión más reciente.

Los usuarios encuentran la fuente más directa de información acerca de pruebas particulares en los catálogos de las editoriales y en el manual que acompaña a cada una. El *Mental Measurement Yearbook* contiene una lista completa de las editoriales especializadas y sus direcciones. Para una fácil referencia, en el apéndice B presentamos los nombres y domicilios de las editoriales cuyas pruebas citamos aquí. Se les puede solicitar los catálogos de pruebas actuales, pero los manuales y las pruebas sólo están a disposición de los usuarios calificados.

El manual de la prueba debe ofrecer la información esencial que se necesita para aplicarla, calificarla y evaluarla; debe incluir instrucciones completas y detalladas, la

clave de calificación, las normas y los datos sobre confiabilidad y validez, y, además, debe informar del número y la naturaleza de las personas en las que se establecieron las normas, la confiabilidad y la validez, así como los métodos utilizados para calcular los índices de estas medidas. En caso de que la información necesaria sea demasiado extensa para el manual, debe dar las referencias al manual técnico o a otros medios impresos en los que se encuentre. En otras palabras, el manual permite que el usuario evalúe la prueba antes de elegirla para sus propósitos particulares. Agreguemos que algunos manuales aún están lejos de esta meta, pero las grandes y más profesionales editoriales especializadas brindan cada vez mayor atención a la preparación de manuales que cumplan los criterios científicos adecuados. Es de esperar que un público ilustrado de usuarios sea la mejor garantía de que dichos criterios se mantendrán y mejorarán.

Es posible encontrar una guía completa, aunque sucinta, para la evaluación de las pruebas psicológicas en los *Standards for Educational and Psychological Testing* ("Criterios para las Pruebas Educativas y Psicológicas") preparados por la Asociación Estadounidense de Psicología (*American Psychological Association*, APA) en colaboración con otras dos asociaciones interesadas en las pruebas, la Asociación Estadounidense de Investigación Educativa (*American Educational Research Association*, AERA) y el Consejo Nacional de Medición Educativa (*National Council on Measurement in Education*, NCME). Publicados inicialmente en 1954, los "Criterios" han sido revisados en 1966, 1974 y 1985; en la actualidad, está en marcha una nueva revisión por parte de las tres asociaciones participantes.

En los ochenta surgió la necesidad de establecer criterios para las pruebas (*Testing Standards*[2]) que no sólo se preocuparan por la calidad técnica de las pruebas sino también por su efecto sobre el bienestar del individuo (véase la figura 1.1, página 30). La naturaleza de la revisión más reciente de los *Testing Standards* indica que esta preocupación es una tendencia en progreso. La figura 1.2 de la página 31 contiene una lista propuesta de criterios preparada por una comisión de las tres asociaciones en 1966. Es evidente que el interés por adaptar la selección de pruebas —así como la preocupación por la interpretación y el uso de sus resultados— al conocimiento sobre las experiencias del examinado muestra un crecimiento continuo. Es digno de observar que toda una sección (Segunda Parte) de la figura 1.2 se titula "Equidad de la prueba". Los usuarios han cobrado conciencia de que la aplicación inadecuada de los instrumentos puede dañar al individuo y disminuir la eficacia de sus contribuciones a la sociedad. Las críticas populares por el mal uso de las pruebas puede haber contribuido a esta mayor conciencia de los examinadores, que a su vez debe disminuir los abusos y, al mismo tiempo, aumentar el reconocimiento público de los beneficios del uso de las pruebas.

[2] Para abreviar, en adelante seguiremos la práctica común de identificarlos como Testing Standards.

CRITERIOS PARA LA APLICACIÓN DE PRUEBAS EDUCATIVAS Y PSICOLÓGICAS

Primera parte. Criterios técnicos para la elaboración y evaluación de los instrumentos

1. Validez
2. Confiabilidad y errores de medición
3. Desarrollo y revisión
4. Escalamiento, normatividad, comparabilidad de las calificaciones e igualdad
5. Publicación: manuales técnicos y guías para el usuario

Segunda parte. Criterios profesionales para el uso de los instrumentos

6. Principios generales de uso
7. Pruebas clínicas
8. Pruebas educativas y aplicación de pruebas psicológicas en las escuelas
9. Uso de pruebas en la consejería
10. Pruebas laborales
11. Licencia y certificación profesional y ocupacional
12. Programa de evaluación

Tercera parte. Criterios para aplicaciones particulares

13. Evaluación de minorías lingüísticas
14. Evaluación de gente con condiciones discapacitantes

Cuarta parte. Criterios para los procedimientos administrativos

15. Aplicación, calificación y reporte
16. Protección de los derechos de los evaluados

Figura 1.1. Temas cubiertos por los *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1985).

LISTA PROPUESTA DE CRITERIOS PARA LOS INSTRUMENTOS EDUCATIVOS Y PSICOLOGICOS

Primera parte. Elaboración, evaluación y documentación de las pruebas

1. Validez
2. Confiabilidad, errores de medición y función informativa de las calificaciones
3. Desarrollo y revisión del instrumento
4. Escalamiento, normas, criterios y comparación de las calificaciones
5. Aplicación, calificación y reporte
6. Documentos de la prueba

Segunda parte. Equidad de las pruebas

7. Equidad y sesgos
8. Protección de los derechos del examinado
9. Aplicación de las pruebas a personas cuya primera lengua no es el inglés
10. Aplicación de las pruebas a individuos con discapacidades

Tercera parte. Aplicaciones de las pruebas

11. Principios generales de uso
12. Pruebas para la evaluación psicológica
13. Pruebas para la evaluación educativa
14. Pruebas laborales, de licencia y certificación
15. Pruebas para los programas de evaluación y política pública

Figura 1.2. Temas elegidos para la edición revisada de los *Standars for Educational and Psychological Testing* (AERA, APA, NCME, 1966). Manuscrito en preparación.

(Reproducido con autorización del Comité Conjunto para los *Standards for Educational and Psychological Testing* [Dianne Brown, directora del proyecto]).

Antecedentes históricos de las pruebas actuales

Una breve revisión de los antecedentes históricos y los orígenes de las pruebas psicológicas nos brindará un panorama útil para comprender las pruebas actuales.[1] La dirección en que éstas han avanzado, así como sus limitaciones y ventajas especiales, podrán entenderse mejor al considerarlas a la luz de sus precursores. En este capítulo nos concretaremos a la revisión de los antecedentes y el desarrollo inicial del movimiento psicométrico como un todo. En los capítulos posteriores analizaremos los desarrollos más recientes relacionados con algunos instrumentos específicos, como los tests de habilidad (capítulos 8 a 12) o los de interés (capítulo 14), así como con las áreas (educativa, industrial, clínica y de consejería) en las que se utilizan tales instrumentos (capítulo 17).

Las raíces de la aplicación de pruebas se pierden en la Antigüedad. Existen relatos del sistema de exámenes del servicio civil que prevaleció en el imperio chino durante 2 000 años (Bowman, 1989). Entre los antiguos griegos, la aplicación de exámenes formaba parte del proceso educativo; las pruebas servían para evaluar el dominio de habilidades físicas e intelectuales (Doyle, 1974). Desde sus inicios en la Edad Media, las universidades europeas basaron los grados y honores en exámenes formales. Con todo, no tenemos que ir más allá del siglo XIX para identificar los

[1] Se encuentra una descripción más detallada de los orígenes de las pruebas psicológicas en F. L. Goodenough (1949) y en J. Peterson (1926). Véase también Boring (1950) y G. Murphy y Kovach (1972) para antecedentes más generales, DuBois (1970) y McReynolds (1975, 1986) para recuentos más recientes de la historia de las pruebas psicológicas, y Anastasi (1965) para los antecedentes históricos del estudio de las diferencias individuales. En Anastasi (1993) hay un repaso general de las tendencias actuales de la psicometría.

principales acontecimientos que forman la base de las pruebas contemporáneas, y a ellos dirigimos ahora nuestra atención.

INTERÉS INICIAL EN LA CLASIFICACIÓN Y LA CAPACITACIÓN DE LAS PERSONAS CON RETARDO MENTAL

El siglo XIX atestiguó el surgimiento del interés por el tratamiento humano de las personas "insanas" y las que sufrían de retardo mental. Hasta ese momento, el destino común de tales individuos era el descuido, el ridículo e incluso la tortura. Con la nueva preocupación por el cuidado adecuado de la gente con problemas mentales, se hizo evidente la necesidad de contar con criterios uniformes para su identificación y clasificación, necesidad que se volvió verdaderamente urgente con la proliferación de instituciones sociales dedicadas a estas personas en todo el mundo. Primero era necesario distinguir entre los individuos insanos y los que sufrían de retardo mental. Los primeros manifestaban trastornos emocionales que podrían o no estar acompañados por un deterioro intelectual a partir de un nivel normal; los segundos se caracterizaban principalmente por una deficiencia intelectual que estaba presente desde el nacimiento o la primera infancia. El primer informe explícito de esta distinción se encuentra quizá en un trabajo publicado en 1838 por el médico francés Esquirol. Se trata de dos volúmenes en los que dedica más de 100 páginas a lo que ahora se conoce como "retardo mental". Esquirol también señala que existen muchos grados de retardo, que varían desde la normalidad hasta la "idiotez profunda". En su esfuerzo por elaborar un sistema para clasificar los diversos grados y variedades de retardo, Esquirol probó varios procedimientos y llegó a la conclusión de que el habla proporciona el criterio más confiable del nivel intelectual del individuo. Es importante decir que los criterios actuales para determinar el retardo mental también son principalmente lingüísticos y que los tests actuales de inteligencia tienen una fuerte carga de contenido verbal. En los capítulos siguientes mostraremos la importante función que cumple la habilidad verbal en nuestro concepto de inteligencia.

De especial significado son las contribuciones de otro médico francés, Seguin, quien fue pionero en la capacitación de los retardados. Luego de rechazar la idea dominante de que el retardo mental era incurable, Seguin (1866/1907) experimentó durante muchos años con lo que llamó el método fisiológico de capacitación, y en 1837 estableció la primera escuela dedicada a la educación de los niños retardados. En 1848 emigró a los Estados Unidos, en donde sus ideas obtuvieron un amplio reconocimiento. Seguin fue el creador de muchas de las técnicas de entrenamiento sensorial y muscular que después fueron adoptadas por las instituciones de asistencia a los retrasados. Con esos métodos, los niños con retardo profundo reciben ejercicio intensivo en discriminación sensorial y para el desarrollo del control motor. Algunos de los procedimientos que estableció Seguin con este propósito fueron luego incorporados a los tests de inteligencia no verbal o de ejecución. Un ejemplo de lo anterior es el

Tablero de Formas* de Seguin (*Seguin Form Board*), en el que el individuo debe insertar bloques de distinta apariencia en los huecos correspondientes tan rápidamente como pueda.

Más de medio siglo después del trabajo de Esquirol y Seguin, el psicólogo francés Alfred Binet recomendaba que los niños que no lograran responder a la educación normal fueran evaluados antes de expulsarlos y que, de ser considerados educables, fueran asignados a cursos de educación especial (T. H. Wolf, 1973). Con sus compañeros de la Sociedad para el Estudio Psicológico del Niño, Binet exhortó al ministro de Instrucción Pública a emprender acciones para mejorar las condiciones de los niños retardados. Un resultado especial fue el establecimiento de una comisión ministerial a cargo de Binet para el estudio de estos infantes. Tal designación fue un acontecimiento trascendental en la historia de las pruebas psicológicas.

LOS PRIMEROS PSICÓLOGOS EXPERIMENTALES

En general, los primeros psicólogos experimentales del siglo XIX no estaban interesados en la medición de las diferencias individuales. Su objetivo principal era la formulación de descripciones generalizadas de la conducta humana, lo que supone que su atención se concentraba en las uniformidades más que en las diferencias conductuales, y que, por lo tanto, las diferencias individuales eran ignoradas o aceptadas como un mal necesario que limitaba la aplicabilidad de las generalizaciones. Así, el hecho de que un individuo reaccionara de manera diferente a otro en las mismas condiciones era considerado como un error, o variabilidad individual, que disminuía la exactitud de las generalizaciones. Ésta era la actitud hacia las diferencias individuales que prevalecía en laboratorios como el fundado por Wundt en Leipzig en 1879, donde se formaron muchos de los primeros psicólogos experimentales.

Al elegir sus áreas de estudio, como en muchas otras fases de su trabajo, los fundadores de la psicología experimental evidenciaban la influencia de sus antecedentes en la fisiología y la física. Los problemas que estudiaban en sus laboratorios tenían que ver principalmente con el tiempo de reacción y con la sensibilidad a estímulos sensoriales como los visuales y los auditivos. Como veremos en otros capítulos, este acento en los fenómenos sensoriales refleja también la naturaleza de las primeras pruebas psicológicas.

Otra influencia que la psicología experimental del siglo XIX ejerció en el curso del movimiento psicométrico se manifiesta en su insistencia por ejercer un control riguroso de las condiciones en las que se realizan las observaciones; por ejemplo, la forma de dar las instrucciones en un experimento de tiempo de reacción puede aumentar o disminuir de modo notable la velocidad de la respuesta, y el brillo o el color del fondo pueden alterar marcadamente la apariencia de un estímulo visual, lo que demuestra la importancia de que todos los participantes del experimento sean observados en condiciones estandarizadas. Esta estandarización del procedimiento llegó a convertirse en una de las señales distintivas de las pruebas psicológicas.

*También conocido como *Tablero de encajamientos*. (N. del T.)

CONTRIBUCIONES DE FRANCIS GALTON

El biólogo inglés Francis Galton fue el principal responsable del inicio del movimiento psicométrico. Un factor común en las numerosas y variadas actividades de investigación de Galton fue su interés por la herencia humana. En el curso de sus investigaciones sobre esta materia, Galton se dio cuenta de la necesidad de medir las características de personas que estaban y no emparentadas, ya que sólo de esta manera podía descubrir, por ejemplo, el grado exacto de parecido entre padres e hijos, hermanos y hermanas, primos o gemelos. Con este propósito, Galton convenció a diversas instituciones educativas de que llevaran registros antropométricos sistemáticos de sus estudiantes. También estableció un laboratorio antropométrico en la Exposición Internacional de 1884 donde, mediante el pago de tres peniques, se medían ciertos rasgos físicos de los visitantes y se los sometía a pruebas de agudeza visual y auditiva, fuerza muscular, tiempo de reacción y otras funciones sensoriomotoras simples. Cuando la exposición cerró, Galton transfirió el laboratorio al Museo de South Kensington, en Londres, donde trabajó durante seis años. Con estos métodos se acumuló gradualmente el primer cuerpo sistemático de datos sobre diferencias individuales en los procesos psicológicos simples.

El propio Galton elaboró varias de las pruebas aplicadas en su laboratorio, muchas de las cuales siguen siendo conocidas, sea en su forma original o modificada. Entre los ejemplos se encuentran la barra de Galton para la discriminación visual de la longitud, el silbato que lleva también su nombre para determinar el mayor tono audible y una serie graduada de pesos para medir la discriminación cinestésica. Galton creía que las pruebas de discriminación sensorial podían servir para estimar el intelecto. A este respecto, había sido también la influencia de las teorías de Locke. Así, Galton escribió: "La única información concerniente a los acontecimientos externos que nos alcanza parece pasar por la avenida de los sentidos, y entre más perceptivos sean los sentidos de las diferencias, mayor será el campo sobre el que pueden actuar nuestro juicio y nuestra inteligencia" (Galton, 1883, p. 27). Galton también se percató de que las personas con retardo mental extremo tienden a mostrar defectos en su capacidad para discriminar entre calor, frío y dolor, una observación que posteriormente fortaleció su convicción de que la capacidad de discriminación sensorial, "en conjunto, sería más elevada entre los intelectualmente aptos" (Galton, 1883, p. 29).

Galton también fue pionero en la aplicación de escalas de calificación y cuestionarios, así como en el uso de la técnica de asociación libre que posteriormente fue utilizada con diversos propósitos. Otra de sus muchas contribuciones se encuentra en la formulación de métodos estadísticos para el análisis de datos sobre las diferencias individuales. Galton tomó y adaptó algunas técnicas matemáticas de forma que pudieran ser utilizadas por investigadores sin conocimientos en la materia que desearan tratar en forma cuantitativa los resultados de sus pruebas. De esta manera extendió considerablemente la aplicación de los procedimientos estadísticos al análisis de datos de las pruebas. Esta fase del trabajo de Galton fue llevada adelante por muchos de sus estudiantes, de los cuales Karl Pearson[2] fue el más eminente.

[2] Para una descripción fascinante de los primeros avances de los principales conceptos estadísticos y de las personas responsables, véase Cowles (1989).

CATTELL Y LOS PRIMEROS TESTS MENTALES

El psicólogo estadounidense James McKeen Cattell ocupa una posición prominente en el adelanto de las pruebas psicológicas. En su trabajo se combinan la recién establecida ciencia de la psicología experimental y el todavía más reciente movimiento psicométrico. Para obtener su doctorado en Leipzig redactó una tesis sobre el tiempo de reacción bajo la dirección de Wundt; y en 1888, mientras enseñaba en Cambridge, su trato con Galton fortaleció su interés por la medición de las diferencias individuales. A su regreso a los Estados Unidos, Cattell participó activamente en el establecimiento de laboratorios de psicología experimental y en la difusión del movimiento psicométrico.

El término "test mental" apareció por primera vez en la bibliografía psicológica en un artículo escrito por Cattell en 1890, que describe una serie de pruebas que cada año se aplicaban a los estudiantes universitarios para determinar su nivel intelectual. Las pruebas, cuya aplicación era individual, incluían mediciones de la fuerza muscular, velocidad de movimiento, sensibilidad al dolor, agudeza visual y auditiva, discriminación de pesos, tiempo de reacción, memoria y cosas similares. En su elección de las pruebas, Cattell compartía la opinión de Galton respecto a la posibilidad de obtener una medida de las funciones intelectuales con el uso de instrumentos de discriminación sensorial y de tiempo de reacción. La preferencia de Cattell por dichas herramientas también se sustentó en el hecho de que le permitían medir con precisión y exactitud las funciones simples, mientras que la obtención de mediciones objetivas de funciones más complejas en ese tiempo parecía una tarea inalcanzable.

Las pruebas de Cattell son características de los instrumentos desarrollados en la última década del siglo XIX, que se aplicaban a escolares, universitarios y adultos. En la Exposición de Columbia realizada en Chicago durante 1893, Jastrow montó una exhibición en la que a los visitantes se los invitaba a realizar pruebas de procesos perceptuales, sensoriales y motores simples y a comparar su habilidad con las normas (J. Peterson, 1926, Philippe, 1894). Los pocos intentos por evaluar estas primeras pruebas arrojaron resultados desalentadores. El desempeño de cada individuo mostraba poca correspondencia de una prueba a otra (Sharp, 1898–1899; Wissler, 1901) y su relación con estimaciones independientes del nivel intelectual basadas en las calificaciones de los maestros (T. L. Bolton, 1891–1892; J. A. Gilbert, 1894) o con las calificaciones académicas (Wissler, 1901) era poca o nula.

Algunas pruebas elaboradas por psicólogos europeos de la época tendían a cubrir funciones algo más complejas. Kraepelin (1895), que estaba interesado sobre todo en la evaluación clínica de pacientes psiquiátricos, preparó una larga serie de pruebas para medir lo que consideraba factores básicos en la caracterización del individuo. Las pruebas, que empleaban principalmente operaciones aritméticas simples, estaban destinadas a medir los efectos de la práctica, la memoria y la susceptibilidad a la fatiga y la distracción. Otro psicólogo alemán, Ebbinghaus (1897), aplicó a escolares pruebas de cálculo aritmético, memoria y completación de oraciones; esta última, que era la más compleja, fue la única que mostró una clara correspondencia con el desempeño escolar.

En un artículo publicado en Francia en 1895, Binet y Henri criticaron el hecho de que casi todas las pruebas disponibles eran sensoriales y se concentraban indebidamente en habilidades especializadas simples. Además, argumentaban que en la medición de las funciones más complejas no se requiere de gran precisión, ya que, en tales funciones, las diferencias individuales son mayores. Propusieron una amplia y variada lista de tests que cubrían funciones como la memoria, la imaginación, la atención, la comprensión, la susceptibilidad a la sugestión, la apreciación estética y muchos otros, en los que podemos reconocer las tendencias que a la postre condujeron al desarrollo de las famosas escalas de inteligencia de Binet.

BINET Y EL SURGIMIENTO DE LOS TESTS DE INTELIGENCIA

Binet y sus colaboradores dedicaron muchos años a la investigación activa e ingeniosa de las formas de medir la inteligencia. Probaron muchos métodos, incluyendo la medición de la forma del cráneo, la cara y la mano, así como el análisis de la escritura; sin embargo, los resultados los llevaron a la convicción de que la medición directa, aunque tosca, de las funciones intelectuales complejas era la más promisoria. Entonces una situación en particular hizo fructificar los esfuerzos de Binet. En 1904, el Ministerio de Educación lo comisionó para que estudiara procedimientos para la educación de niños retardados. Fue en relación con los objetivos de esta comisión que Binet preparó, en colaboración con Simon, la primera escala de Binet-Simon (Binet y Simon, 1905).

Esta escala, conocida como la escala de 1905, constaba de 30 problemas o tests arreglados en orden de dificultad creciente. El nivel de dificultad se estableció empíricamente aplicándolos a 50 niños normales de tres a 11 años y a algunos niños y adultos retardados. Las pruebas fueron diseñadas para cubrir una amplia variedad de funciones, con énfasis especial en el juicio, la comprensión y el razonamiento, que Binet consideraba los componentes principales de la inteligencia. Aunque incluyeron pruebas sensoriales y perceptuales, en esta escala se encuentra una proporción de contenido verbal mayor que en la generalidad de las pruebas de la época. Como la escala de 1905 fue presentada como un instrumento preliminar y tentativo, no se formuló un método objetivo preciso para obtener una puntuación total.

En la segunda escala, la de 1908, aumentó el número de tests, se eliminaron algunos de la primera que resultaron insatisfactorios y todos fueron agrupados en niveles de edad sobre la base del desempeño de alrededor de 300 niños normales de entre tres y 13 años. De este modo, en el nivel de tres años se ubicaron todos los tests que pasaban del 80 al 90 por ciento de los niños normales de tres años; en el nivel de cuatro años, los que aprobó el mismo porcentaje de niños normales de esa edad; y así sucesivamente hasta los 13. La calificación del niño en toda la prueba podía entonces expresarse como el *nivel mental* correspondiente a la edad de los niños normales cuya ejecución había igualado. En las diversas traducciones y adaptaciones de las escalas de Binet el término de "nivel mental" fue sustituido por el de "edad mental", cuya fácil comprensión indudablemente contribuyó a popularizar los tests de

inteligencia;[3] sin embargo, el propio Binet evitaba el uso del término por sus implicaciones no verificadas de desarrollo y prefería el más neutral de "nivel mental" (T. H. Wolf, 1973).

En 1911, año en que Binet murió intempestivamente, apareció la tercera revisión de la Escala de Binet-Simon, que no presenta mayores modificaciones, salvo revisiones menores, cambios en la localización de algunos tests, la adición de otros en varios niveles de edad y la extensión de la escala al nivel adulto.

Incluso antes de la revisión de 1908, los tests de Binet-Simon atrajeron la atención de los psicólogos de todo el mundo. En muchos países, aparecieron traducciones y adaptaciones, pero la primera fue la de H. H. Goddard, en ese entonces psicólogo investigador en la Escuela de Capacitación de Vineland para niños con retardo mental. La revisión de Goddard resultó clave para que la profesión médica aceptara los tests de inteligencia (Zenderland, 1987). Apareció en un momento propicio para satisfacer la urgente necesidad de un instrumento estandarizado para diagnosticar y clasificar a las personas con retardo mental. No obstante, como herramienta de evaluación pronto fue dejada atrás por el Stanford-Binet, un instrumento más extenso y psicométricamente refinado, elaborado por L. M. Terman y sus colaboradores en la Universidad de Stanford (Terman, 1916). Este instrumento utilizó por primera vez el cociente de inteligencia (CI) o razón entre la edad mental y la cronológica. La última revisión se emplea extensamente y la estudiaremos de manera más completa en el capítulo 8. También resulta de especial interés la primera revisión Kuhlmann-Binet, que amplió la escala hasta la edad de tres meses (Kuhlmann, 1912) y representa uno de los primeros esfuerzos por elaborar tests de inteligencia para preescolares e infantes.

PRUEBAS COLECTIVAS

Los tests de Binet, así como todas sus revisiones, son *escalas individuales* en el sentido de que sólo pueden aplicarse a una persona y luego a otra. Muchos de los tests de esas escalas requieren respuestas orales del examinado o la manipulación de materiales. Algunos exigen tomar el tiempo de las respuestas de cada individuo. Por éstas y otras razones, las pruebas no pueden adaptarse a la aplicación colectiva. Otra característica de los tests de Binet es que requieren de un examinador muy capacitado, pues se trata de instrumentos esencialmente clínicos adecuados para el estudio intensivo de casos individuales.

Las *pruebas de grupo* similares a la primera escala de Binet fueron elaboradas para satisfacer una necesidad práctica. Cuando los Estados Unidos ingresaron en 1917 en la Primera Guerra Mundial, la Asociación Estadounidense de Psicología formó una comisión para considerar cómo podía contribuir la psicología a la conducción de la

[3] F. L. Goodenough (1949, pp. 50–51) observó que en 1887, 21 años antes de la aparición de la Escala Binet-Simon de 1908, S. E. Chaille publicó en el *New Orleans Medical and Surgical Journal* una serie de pruebas para infantes, arregladas de acuerdo con la edad en que solían ser pasadas. Debido en parte a la limitada circulación de la revista y quizá también a que la comunidad científica no estaba preparada, en esa época pasó inadvertido el significado del concepto de escala de edad. La propia escala de Binet recibió la influencia del trabajo de algunos de sus contemporáneos, principalmente Blin y Damaye, que prepararon un conjunto de preguntas orales de las que obtenían una sola calificación global para cada niño (T. H. Wolf, 1973).

guerra. La comisión, dirigida por Robert M. Yerkes, reconoció la necesidad de la rápida clasificación del nivel intelectual general del millón y medio de reclutas. Esta información era importante para tomar muchas decisiones administrativas, incluyendo el rechazo o la dispensa del servicio militar, la asignación a diferentes servicios o la admisión a campos de entrenamiento de oficiales. En este ambiente se formuló el primer test colectivo de inteligencia. Para realizar la tarea, los psicólogos acudieron a todos los materiales disponibles, especialmente a un test colectivo de inteligencia no publicado, preparado por Arthur S. Otis, que cedió al ejército. Una contribución importante de este instrumento, elaborado por Otis cuando era estudiante en uno de los cursos de grado de Terman, fue la introducción de preguntas de opción múltiple y otros reactivos "objetivos".

Las pruebas que finalmente desarrollaron los psicólogos del ejército podían aplicarse a grupos grandes y llegaron a conocerse como Army Alpha y Army Beta (pruebas Alfa y Beta). La primera se ocupaba de las pruebas generales de rutina, mientras que la segunda era una escala no lingüística utilizada con reclutas iletrados o procedentes de países cuyo idioma no era el inglés.

Poco después de terminar la Primera Guerra Mundial, las pruebas del ejército fueron liberadas para que se aplicaran a civiles; además de pasar por muchas revisiones, las pruebas Alfa y Beta sirvieron como modelo para la mayor parte de los tests colectivos de inteligencia, lo que contribuyó al notable crecimiento del movimiento psicométrico. Muy pronto se prepararon tests colectivos de inteligencia para todas las edades y personas, desde preescolares hasta universitarios. Los programas de evaluación a gran escala, otrora imposibles, se emprendieron con gran entusiasmo. Como las pruebas colectivas fueron diseñadas como instrumentos de evaluación masiva, no sólo permitieron la medición simultánea de grandes grupos, sino que también simplificaron las instrucciones y los procedimientos de aplicación, con lo que disminuyó la capacitación requerida del aplicador. Los maestros empezaron a aplicar en sus grupos los tests de inteligencia y los estudiantes universitarios eran evaluados de rutina antes de su ingreso; se iniciaron amplios estudios de grupos especiales de adultos, como los prisioneros, y muy pronto el público general cobró conciencia del CI.

La aplicación de esos tests colectivos excedió a su mejoramiento técnico. En la prisa por obtener puntuaciones y sacar conclusiones prácticas de los resultados, a menudo se olvidaba que las pruebas aún eran técnicamente imperfectas, de modo que cuando no lograban cumplir las injustificadas expectativas, se generaba escepticismo y hostilidad hacia todas las pruebas. De este modo, el auge psicométrico de los años veinte, basado en el uso indiscriminado de los instrumentos, puede haber demorado el progreso de las pruebas psicológicas.

TESTS DE APTITUD

Aunque al principio los tests de inteligencia fueron diseñados para cubrir una amplia variedad de funciones que permitieran estimar el nivel general de inteligencia del individuo, pronto se hizo evidente que su alcance era muy limitado. No todas las funciones importantes estaban representadas. De hecho, la mayor parte de los tests de inteligencia eran principalmente mediciones de la habilidad verbal y, en menor grado, de la

habilidad para manejar relaciones numéricas y otras relaciones abstractas y simbólicas. Gradualmente, los psicólogos reconocieron que el término "test de inteligencia" era un nombre inadecuado, ya que sólo medían ciertos aspectos de ésta.

Para que resultaran seguros, los instrumentos tocaban habilidades que son de gran importancia en la cultura para la que fueron diseñados, pero pronto se reconoció la conveniencia de contar con designaciones más precisas en términos de la información que podían ofrecer; por ejemplo, ahora se conoce como tests de aptitud escolar a algunos instrumentos que en los veinte habrían sido llamados de inteligencia. Este cambio en la terminología procede del hecho de que muchos de los llamados tests de inteligencia miden la combinación de habilidades requeridas y fomentadas por el trabajo universitario.

Incluso antes de la Primera Guerra Mundial, los psicólogos empezaban a admitir la necesidad de contar con tests de aptitudes especiales que complementaran los de inteligencia global. En especial, se prepararon *pruebas de aptitudes especiales* para uso en la orientación vocacional y en la selección y clasificación de personal industrial y militar. Entre las más empleadas están los tests de aptitud mecánica, profesional, musical y artística.

La evaluación crítica de las pruebas de inteligencia que siguió al uso indiscriminado durante los veinte también reveló el hecho notable de que el desempeño del individuo a menudo mostraba una marcada variación en diferentes partes de la prueba. Esto resultó en especial evidente en las pruebas colectivas, en las que los reactivos suelen separarse en subpruebas de contenido relativamente homogéneo; por ejemplo, una persona podía obtener una puntuación más bien alta en la subprueba verbal y un pobre resultado en la numérica o viceversa. Dicha variabilidad interna es hasta cierto grado discernible en un test como el Stanford-Binet en el que, por ejemplo, todos los reactivos que incluyen palabras pueden resultar difíciles para cierto individuo, mientras que los que emplean imágenes o diagramas geométricos pueden resultarle ventajosos.

Los usuarios, y sobre todo los clínicos, a menudo se valían de esas comparaciones internas para obtener un conocimiento más profundo de la estructura psicológica del individuo. De este modo, al evaluar un caso individual no sólo se examinaba el CI o algún otro resultado global, sino también el desempeño en grupos de reactivos o subtests; sin embargo, en general no se recomienda esta práctica, porque los tests de inteligencia no fueron diseñados con el propósito de realizar un análisis diferencial de las aptitudes. Con frecuencia, los subtests comparados contienen muy pocos reactivos para producir una estimación estable o confiable de determinada habilidad. Entonces, la diferencia de las puntuaciones de los subtests se puede revertir si vuelve a examinarse al individuo otro día o con otra forma del mismo instrumento. Para realizar esas comparaciones en el mismo sujeto, es necesario diseñar las pruebas con el objeto de que revelen diferencias de ejecución en diversas funciones.

Al mismo tiempo que la aplicación práctica demostraba la necesidad de contar con múltiples tests de aptitud, un desarrollo paralelo en el estudio de la organización de los rasgos iba proporcionando los medios para elaborarlos. Los estudios estadísticos sobre la naturaleza de la inteligencia habían explorado las relaciones entre las puntuaciones obtenidas por muchas personas en una gran variedad de pruebas. El psicólogo inglés

Charles Spearman (1904, 1927) inició estas investigaciones durante la primera década del siglo XX. Los desarrollos metodológicos subsecuentes, basados en el trabajo de estudiosos ingleses y estadounidenses, como T. L. Kelley (1928) y L. L. Thurstone (1938, 1947b), llegaron a conocerse como *análisis factorial*.

En el capítulo 11 examinaremos de manera exhaustiva las contribuciones de los métodos del análisis factorial en la elaboración de pruebas. Por ahora basta con destacar que los datos obtenidos con estos procedimientos señalaron la presencia de diversos factores o rasgos relativamente independientes, algunos representados en diversas proporciones en los tests tradicionales de inteligencia, como, por ejemplo, la comprensión verbal y el razonamiento numérico; otros, como la aptitud espacial, perceptual y mecánica, se encuentran más a menudo en los instrumentos para la medición de aptitudes especiales que en los de inteligencia.

Uno de los principales resultados prácticos del análisis factorial fue el desarrollo de las *baterías de aptitudes múltiples*, diseñadas para proporcionar una medida de la posición del individuo en una serie de rasgos. En lugar de una puntuación total, o CI, estos instrumentos obtienen una puntuación aparte para rasgos como la comprensión verbal, la aptitud numérica, la visualización espacial, el razonamiento aritmético y la rapidez perceptual, lo que permite realizar el análisis intrasujeto o diagnóstico diferencial en el mismo sujeto que los usuarios trataron de obtener por muchos años con los resultados crudos y a menudo erróneos de los tests de inteligencia. Además, dado que cubren algunos de los rasgos que generalmente no se incluían en ese tipo de instrumentos también incorporaron en un programa amplio y sistemático de evaluación gran parte de la información que antes se obtenía de las pruebas de aptitud especial.

Las baterías de aptitudes múltiples representan un desarrollo relativamente tardío en el campo de las pruebas, ya que casi todas aparecieron a partir de 1945. A este respecto es de destacar el trabajo de los psicólogos militares durante la Segunda Guerra Mundial. Gran parte de la investigación conducida en las fuerzas armadas se basó en el análisis de factores y se orientó a la preparación de baterías de aptitudes múltiples; por ejemplo, en la fuerza aérea se construyeron baterías especiales para pilotos, bombarderos, operadores de radio, rastreadores y muchos otros especialistas. El informe de las baterías que elaboró la fuerza aérea ocupa al menos nueve de los 19 volúmenes dedicados al programa psicológico de la aviación durante la Segunda Guerra Mundial (*Army Air Forces*, 1947–1948). La investigación en esas líneas aún sigue en progreso con el patrocinio de varias ramas de las fuerzas armadas. También se han realizado baterías para uso civil y se aplican lo mismo en la orientación vocacional y educativa que en la selección y clasificación de personal. En los capítulos 10 y 17 presentaremos algunos ejemplos.

Un adelanto más reciente, surgido a finales de los ochenta y comienzos de los noventa, permite la integración de dos métodos al principio contradictorios de medición mental representados por los tests tradicionales de inteligencia y las baterías de aptitudes múltiples (Anastasi, 1994). En la actualidad, se reconoce que la habilidad del ser humano puede ser evaluada de manera adecuada a diferentes niveles de profundidad, desde las aptitudes muy definidas de las pruebas específicas (o incluso de reactivos), pasando por el nivel de los rasgos hasta una puntuación global como el tradicional CI. Para cada propósito de evaluación hay un nivel apropiado de profundidad. En

correspondencia, los tests de inteligencia que se han desarrollado recientemente, como las Escalas de Habilidad Diferencial o las revisiones recientes de tests anteriores, como la cuarta edición del Stanford-Binet (que explicaremos en el capítulo 8), combinan la amplia cobertura de diversas aptitudes con calificaciones flexibles de niveles múltiples para propósitos concretos de evaluación. Aunque se trata de dos ejemplos de tests de inteligencia de aplicación individual, el mismo método amplio y flexible de preparación y uso de los instrumentos ha tenido un impacto en las baterías de aplicación colectiva, como las que veremos en el capítulo 10. En el capítulo 11 estudiaremos la base teórica y las implicaciones prácticas de esta combinación de tests de habilidad en relación con los adelantos recientes, relativos a la naturaleza de la inteligencia.

PRUEBAS ESTANDARIZADAS DE APROVECHAMIENTO

Mientras los psicólogos se ocupaban de los tests de inteligencia y de aptitud, los exámenes escolares tradicionales experimentaban avances técnicos (O. W. Caldwell y Courtis, 1923; Ebel y Damrin, 1960). Un paso importante en esta dirección lo dieron las escuelas públicas de Boston en 1845 cuando los exámenes escritos sustituyeron al interrogatorio oral de los estudiantes por examinadores visitantes. Entre los argumentos ofrecidos en apoyo de esta innovación estaba que los exámenes escritos ponían a los estudiantes en una situación uniforme, permitían una cobertura más amplia del contenido, reducían el elemento azaroso en la elección de reactivos y eliminaban la posibilidad de favoritismo por parte del examinador. Todos estos argumentos tienen un sonido familiar: fueron utilizados mucho después para justificar la sustitución de los reactivos de ensayo por los reactivos objetivos de opción múltiple.

A la vuelta del siglo empezaron a aparecer las primeras pruebas estandarizadas para medir los resultados de la instrucción escolar. Encabezadas por el trabajo de E. L. Thorndike, las pruebas empleaban principios de medición tomados del laboratorio psicológico. Entre los ejemplos se incluyen escalas para calificar la calidad de la escritura y la redacción, así como pruebas de ortografía, aritmética, cálculo y razonamiento aritmético. Después llegaron las baterías de aprovechamiento, iniciadas con la publicación de la primera edición de la Prueba de Aprovechamiento de Stanford (*Stanford Achievement Test*) en 1923. Sus autores fueron tres líderes en la elaboración de pruebas: Truman L. Kelley, Giles M. Ruch y Lewis M. Terman. Presagiando muchas características de la psicometría moderna, esta batería proporcionó medidas comparables de ejecución en diferentes materias escolares, evaluadas en términos de un solo grupo normativo.

Al mismo tiempo se acumulaban evidencias de la falta de acuerdo entre los maestros al calificar las pruebas de ensayo. Para 1930 era ampliamente reconocido que las pruebas de ensayo no sólo requerían más tiempo para los examinadores y examinados, sino que también arrojaban resultados menos confiables que el "nuevo tipo" de reactivos objetivos.[4] En la medida en que estos últimos llevaron al uso creciente de las pruebas estandarizadas de aprovechamiento, hubo un énfasis mayor en el diseño

[4] La investigación relacionada con la relativa eficacia de los reactivos de ensayo y objetivos se trata en el capítulo 17, en la parte que aborda la utilización educativa de las pruebas

de reactivos para probar la comprensión y la aplicación del conocimiento y otros objetivos educativos más amplios. La década de los treinta también presenció la introducción de las máquinas para calificar exámenes a las cuales podían adaptarse fácilmente las nuevas pruebas objetivas.

Otro desarrollo digno de mención fue el establecimiento en los Estados Unidos de programas estatales, regionales y nacionales de evaluación. Probablemente el más conocido sea el de la Junta Universitaria de Exámenes de Admisión a la Universidad (*College Entrance Examination Board*, CEEB). Establecido a principios de siglo para reducir la duplicación de los exámenes de admisión presentados por los estudiantes de reciente ingreso a la universidad, el programa ha experimentado cambios profundos en sus procedimientos de evaluación y en el número y la naturaleza de las universidades participantes, cambios que reflejan los adelantos tanto en la aplicación de pruebas como en la educación. En 1947, las funciones de evaluación del CEEB se combinaron con las de la Corporación Carnegie y las del Consejo Estadounidense de Educación para formar el Servicio de Pruebas Educativas (*Educational Testing Service*, ETS). En los años posteriores, el ETS asumió la responsabilidad de un número creciente de programas de evaluación al servicio de universidades, escuelas profesionales, dependencias gubernamentales y otras instituciones. Debe hacerse mención del Programa de Evaluación de Universidades Estadounidenses (*American College Testing Program*), establecido en 1959 para seleccionar a los aspirantes a las universidades no incluidas en el programa CEEB y de varios programas nacionales de evaluación para premiar a los estudiantes talentosos.

Las pruebas de aprovechamiento no sólo se utilizan con propósitos educativos, sino también en la selección de solicitantes de empleos en la industria y el gobierno. Ya citamos el uso sistemático en el servicio civil de exámenes en el imperio chino desde aproximadamente el año 150 a.C. (Bowman, 1989). En los tiempos modernos, la selección de empleados gubernamentales por medio de exámenes fue introducida en los países europeos a finales del siglo XVIII y comienzos del XIX. En 1883, la Comisión del Servicio Civil de los Estados Unidos (*U.S. Civil Service Commission*) estableció el uso de exámenes competitivos como procedimiento regular (Kavruck, 1956). Las técnicas de elaboración de pruebas creadas durante y antes de la Primera Guerra Mundial fueron introducidas en el programa de evaluación del Servicio Civil de los Estados Unidos con el nombramiento de L. J. O'Rourke como director de la recién establecida división de investigación en 1922. En la actualidad, este trabajo lo realiza un sofisticado equipo de investigación de la Oficina de Administración de Personal de los Estados Unidos (*U.S. Office of Personnel Management*).

En la medida en que crecía la participación de psicólogos con preparación en psicometría en la formulación de pruebas estandarizadas de aprovechamiento aumentaba la semejanza de sus aspectos técnicos con el de los tests de inteligencia y de aptitud. Los procedimientos para la elaboración y evaluación de todas esas nuevas pruebas tenían mucho en común. Los crecientes esfuerzos por preparar pruebas de aprovechamiento que midieran la consecución de amplias metas educativas, en contraste con el recuerdo de hechos triviales, también hizo que el contenido de las pruebas de aprovechamiento se asemejara al de los tests de inteligencia. En la actualidad, la diferencia entre los dos instrumentos corresponde principalmente al grado de especificidad del contenido y el grado en que el instrumento presupone una instrucción previa.

EVALUACIÓN DE LA PERSONALIDAD

Otra área de interés de las pruebas psicológicas son los aspectos afectivos o no intelectuales de la conducta, los que revisaremos del capítulo 13 al 16. Los instrumentos diseñados con este propósito suelen conocerse como tests de personalidad, aunque muchos psicólogos prefieren emplear el término "personalidad" en un sentido más amplio para referirse al individuo en su totalidad. De acuerdo con esto, tanto los rasgos intelectuales como los no intelectuales deberían agruparse bajo dicho rubro; sin embargo, en la terminología psicométrica es más común el uso de la expresión "test de personalidad" para referirse a la medición de características como los estados emocionales, las relaciones interpersonales, la motivación, los intereses y las actitudes.

El uso que hizo Kraepelin de la prueba de asociación libre con pacientes psiquiátricos es un antecedente de los tests de personalidad. En esta prueba, se presentan al examinado palabras estímulo especialmente seleccionadas y se le pide que responda a cada una con la primera palabra que le venga a la mente. Kraepelin (1892) también utilizó esta técnica para estudiar los efectos psicológicos de la fatiga, el hambre y las drogas, y concluyó que todos esos agentes incrementan la frecuencia relativa de asociaciones superficiales. Sommer (1894), que también trabajó en la última década del siglo XIX, sugirió que la prueba de asociación libre podría utilizarse para diferenciar formas de trastorno mental. Después, la técnica ha sido utilizada con distintos propósitos de evaluación y aún sigue empleándose. Debe mencionarse el trabajo de Galton, Pearson y Cattell en la preparación de cuestionarios estandarizados y escalas de calificación. Aunque originalmente estaban destinados a otros propósitos, estos procedimientos fueron utilizados para elaborar algunos de los tests de personalidad que hoy son más comunes.

El prototipo de cuestionario de personalidad, o *inventario autodescriptivo* (capítulo 13), es la Hoja de Datos Personales (*Personal Data Sheet*) creada por Woodworth durante la Primera Guerra Mundial (DuBois, 1970; Franz, 1919, pp. 171–176; L. R. Goldberg, 1971; Symonds, 1931, capítulo 5). El cuestionario fue diseñado como una herramienta de selección para identificar a los individuos gravemente perturbados que deberían ser excluidos del servicio militar. El cuestionario constaba de una serie de preguntas que versaban sobre síntomas psicopatológicos comunes y en las que los individuos respondían sobre sí mismos. Se obtenía una puntuación total, contando el número de síntomas indicados. Este instrumento no se terminó y no pudo emplearse a tiempo, antes de que terminara la guerra, pero inmediatamente después se prepararon formas para uso civil, incluyendo una forma especial para aplicar a niños. Más aún, la Hoja de Datos Personales de Woodworth sirvió como modelo para la mayor parte de los inventarios de ajuste emocional. En algunos de estos cuestionarios se hacía un intento por subdividir el ajuste emocional en formas más específicas, como el ajuste al hogar, el escolar y el vocacional. Otros instrumentos se concentraban en un área más estrecha de conducta o en respuestas más claramente sociales, como las de dominancia-sumisión en el trato personal. El último avance fue la elaboración de instrumentos para cuantificar la expresión de actitudes e intereses (capítulo 14), que también se basaban principalmente en las técnicas de cuestionario.

Otro método de medición de la personalidad se encuentra en la aplicación de *tests situacionales* y de *ejecución* (capítulo 16), en las que el examinado debe realizar una

tarea cuyo propósito a menudo está encubierto. La mayor parte simula con mucho realismo situaciones de la vida cotidiana. La primera aplicación extensa de las técnicas está en las pruebas elaboradas por Hartshorne, May y colaboradores a finales de los veinte y principios de los treinta (1928, 1929, 1930). Esta serie, estandarizada para escolares, se interesaba en conductas tales como copiar, mentir, robar, cooperar y persistir. Era posible obtener puntuaciones cuantitativas objetivas en numerosos instrumentos específicos. Otro ejemplo, éste para los adultos, se encuentra en la serie de tests situacionales preparados durante la Segunda Guerra Mundial por el Programa de Evaluación de la Oficina de Servicios Estratégicos (*Office of Strategic Services*, OSS, 1948). Estos tests se interesaban en la conducta emocional y social sutil y relativamente compleja y su aplicación requería de condiciones más bien elaboradas y de personal capacitado, además de que la interpretación de las respuestas era relativamente subjetiva.

Las *técnicas proyectivas* (capítulo 15) representan el tercer método de estudio de la personalidad que ha mostrado un crecimiento notable, en especial entre los clínicos. Estos instrumentos presentan al cliente un estímulo no muy estructurado, lo que permite una considerable libertad en su solución. La suposición que fundamenta este método es que el individuo proyectará mediante el estímulo su estilo característico de respuesta. Como en los tests situacionales o de ejecución, el propósito de las técnicas proyectivas está más o menos encubierto, lo que reduce la posibilidad de que el individuo cree deliberadamente una impresión deseable. La prueba de asociación libre, que ya citamos, es una de las primeras técnicas proyectivas. Los tests de frases incompletas también se han utilizado de esta manera. Otras tareas que aparecen en las técnicas proyectivas incluyen el dibujo, el arreglo de juguetes para crear una escena, la dramatización extemporánea y la interpretación de manchas de tinta.

Todos los tests de personalidad disponibles presentan ciertas dificultades prácticas y teóricas. Cada método tiene sus propias ventajas y desventajas. En conjunto, los tests de personalidad han quedado detrás de los de habilidad en cuanto a logros prácticos, pero esa falta de progreso no puede atribuirse a un esfuerzo insuficiente. La investigación sobre la medición de la personalidad ha alcanzado proporciones impresionantes desde 1950, y muchos instrumentos ingeniosos y mejoras técnicas están en investigación. Lo que explica el lento avance en el área son más bien las dificultades específicas que se encuentran en la medición de la personalidad.

A partir de la investigación actual con los tests de personalidad están surgiendo dos tendencias unificadoras importantes (véase Anastasi, 1985b, 1992a, 1993; Digman, 1990; L. R. Goldberg, 1993; Simon, 1994). Primera, cada vez hay más evidencias de la influencia recíproca de los rasgos afectivos (de "personalidad") y cognoscitivos (de "habilidad") tanto en el desempeño de tareas como en el desarrollo conductual. Se ha llegado a la conclusión de que la distinción tradicional entre los dos rasgos ha sido impuesta artificialmente por razones de conveniencia en la descripción y medición de diferentes aspectos de la conducta. Segunda, el análisis teórico de la naturaleza y composición de la personalidad apoya la integración de los rasgos cognoscitivos y afectivos en un modelo amplio de la actividad humana que incluye todas las formas de conducta. Este modelo relaciona la investigación básica de los rasgos intelectuales (capítulo 11) y los afectivos (capítulo 13).