INTRODUCCIÓN A LA PSICOMETRÍA

SILVIA TORNIMBENI
EDGARDO PÉREZ
FABIÁN OLAZ
(compiladores)

Con la colaboración de
NURIA CORTADA DE KOHAN
ALBERTO FERNÁNDEZ
MARCOS CUPANI

INTRODUCCIÓN A LA PSICOMETRÍA

PAIDÓS
Buenos Aires
Barcelona
México

Introducción a la psicometría / Silvia Tornimbeni...[et.al.]. - 1a ed. - Buenos Aires : Paidós, 2008.
288 p. ; 22x16 cm. (Evaluación Psicológica; 21085)

ISBN 978-950-12-6085-4

1. Psicología. 2. Psicometría.
CDD 153.9

Cubierta de Gustavo Macri

1ª edición, 2008

Queda hecho el depósito que previene la ley 11.723
Impreso en la Argentina. Printed in Argentina

Impreso en Primera Clase, California 1231, Ciudad de Buenos Aires
en febrero de 2008

Tirada: 3.000 ejemplares

ISBN: 978-950-12-6085-4

ÍNDICE

LOS AUTORES

SILVIA TORNIMBENI

Licenciada en Psicología, especialista en Psicometría y Psicología Educacional. Profesora titular de la cátedra Técnicas y Psicométricas y miembro del Consejo Directivo, Facultad de Psicología, Universidad Nacional de Córdoba. Posee una extensa trayectoria en gestión, evaluación y formación de recursos humanos. Autora de libros, capítulos de libros y numerosas publicaciones académicas. Investigadora en el Programa de Incentivos SECyT.

EDGARDO PÉREZ

Doctor en Psicología, especialista en Psicometría y Desarrollo de Carrera. Profesor adjunto de la cátedra Técnicas Psicométricas y miembro del Comité Académico de la Carrera de Doctorado, Facultad de Psicología, Universidad Nacional de Córdoba. Autor de libros, capítulos de libros y artículos en revistas nacionales e internacionales. Director de tesis de grado y posgrado, así como de becas de investigación SECyTy Conicet. Investigador en el Programa de Incentivos de SECyT.

FABIÁN OLAZ

Licenciado en Psicología, especialista en Psicometría y Teoría Social Cognitiva. Becario de Conicet. Profesor de la cátedra Técnicas Psicométricas, Facultad de Psicología, Universidad Siglo XXI. Autor de libros, capítulos de libros y artículos en revistas nacionales e internacionales.

PRÓLOGO

Este libro es una introducción a la psicometría en un doble sentido: por un lado, permite adentrarse en los temas e instrumentos clásicos de la disciplina y, por otro, ofrece un panorama de los desarrollos más actuales, tales como los tests informatizados.

Se tratan aquí los temas básicos relacionados con la confiabilidad y la validez, y otros más avanzados como el metá-analisis y el análisis factorial. Este último método, fundamental en el desarrollo actual de la psicometría, es objeto de una presentación básica pero reflexiva y completa, que asegura la comprensión de sus objetivos, fundamentos y aplicaciones. Con respecto a los instrumentos de medición psicológica, se analizan pruebas tradicionales (por ejemplo, los tests de inteligencia) y también otras concebidas para la evaluación de constructos de reciente interés, como las creencias de autoeficacia.

La atención a los métodos multivariables, de inmensa importancia hoy, es también una característica de este libro. A la consideración del análisis factorial, ya mencionada, se agrega una presentación del análisis de regresión múltiple, de especial significación en los estudios de validez.

El libro contiene exposiciones accesibles, con énfasis en los fundamentos conceptuales. Los temas son abordados de manera de hacer posibles las aplicaciones por parte del lector. Esta preocupación impregna toda la obra y merece también secciones especiales como las dedicadas a la adaptación de tests a otras culturas (de enorme importancia en países como el nuestro) y a procedimientos informáticos de especial interés para la psicometría contemporánea.

Celebro el lugar que las consideraciones teóricas tienen en este libro. Así, por ejemplo, se bosquejan diversas concepciones actuales sobre la estructura de la inteligencia y otros planteamientos que han tenido importantes aplicaciones en psicometría, como la teoría social-cognitiva de Bandura. A la presentación de la concepción clásica de los tests sigue una exposición sintética de la teoría de respuesta al ítem, de especial interés y significación actuales.

Creo que el texto resultará valioso para el estudiante y también para quienes se interesen por las cuestiones metodológicas y epistemológicas relativas a la posibilidades de matematización en las ciencias del comportamiento. Adicionalmente, ofrece al estudioso un panorama de los aportes al desarrollo y la adaptación de pruebas psicológicas realizados en nuestro país, generalmente poco conocidos.

Prologo este libro con profunda satisfacción. Los textos sobre la materia en idioma inglés son abundantes. Esta obra, fruto de la labor de investigadores y docentes de la Univesidad Nacional de Córdoba, extiende la bibliografía en idioma castellano y la enriquece de manera apreciable.

PROFESOR LIVIO GRASSO
Facultad de Psicología
Universidad Nacional de Córdoba

PREFACIO

"El desarrollo de nuestra ciencia seguirá probablemente el modelo de toda ciencia, haciéndose cada vez más matemática a medida que las ideas fundamentales se formulen de manera más rigurosa."

L. L. THURSTONE

Hace algunos años realizamos una encuesta para investigar el uso de tests por parte de psicólogos y psicopedagogos de la ciudad de Córdoba, Argentina (Pérez y Gay, 1991). Una de las principales conclusiones de ese estudio fue que existía una necesidad imperiosa de especialistas capacitados para construir tests en base a las necesidades de nuestra comunidad, así como para adaptar aquéllos elaborados en otros países. También pudimos identificar que la mayoría de los profesionales entrevistados realizaban un uso técnicamente inadecuado de los tests. Más recientemente, Fernández, Marino, Villacorta y Pérez (2000) replicaron esta investigación con resultados similares. En efecto, solamente la mitad de los encuestados informó utilizar tests en sus actividades profesionales y un elevado porcentaje de los entrevistados no otorgaba la importancia debida a los requisitos técnicos y fundamentos teóricos de las pruebas psicométricas que utilizaban. Para evitar este empleo inapropiado de los tests es esencial que los estudiantes de psicología y carreras afines, así como los profesionales usuarios de pruebas psicológicas, adquieran y/o incrementen su formación en la teoría y técnica de los tests.

Este libro intenta realizar una contribución acotada en ese sentido, como herramienta básica de consulta en nuestra re-

gión, y está pensado para un lector sin conocimientos psicométricos aunque con formación básica en estadística aplicada a la ciencia del comportamiento. En efecto, la psicometría es en gran medida una estadística aplicada y, por consiguiente, requiere la comprensión de términos básicos tales como medidas de tendencia central y dispersión, niveles de medición, correlación entre variables, curva normal y probabilidad, entre otros. No obstante, algunos de estas nociones se repasan sintéticamente en los diferentes capítulos de este texto.

El investigador dedicado a la construcción, adaptación de tests o a la investigación psicométrica puede utilizar esta publicación como material inicial que le facilite la consulta posterior a textos más especializados, algunos de los cuales mencionamos repetidamente en este manual (Anastasi y Urbina, 1998; Cronbach, 1998; Hogan, 2004; Kline, 2000; Muñiz, 2001; Aiken, 2003; Martínez Arias, 1995; APA, 1999, por ejemplo). Otra característica distintiva de este manual introductorio es la mención continua de los tests desarrollados o adaptados en el país. Esta información debería ser de utilidad para el estudioso de la psicometría o el usuario de tests, quienes muchas veces desconocen la producción local en esta disciplina.

La primera parte del volumen aborda la controvertida problemática de la medición en psicología. En esta primera sección también se describe sintéticamente la evolución histórica de los tests, así como una propuesta de clasificación de estos instrumentos de medición, incluyendo algunas referencias al trabajo realizado en la Argentina para construir y adaptar tests. La segunda parte desarrolla nociones fundamentales relacionadas con los requisitos técnicos que deben reunir los tests para su uso en situaciones reales de evaluación de personas: confiabilidad, validez, interpretación de las puntuaciones, construcción y adaptación de tests, en ese orden. Finalmente, en la tercera parte se revisan las teorías de los tests psicológicos: la teoría clásica de los tests y la teoría de respuesta al ítem. Un apéndice complementa el texto, ilustrando algunos análisis psicométricos esenciales mediante el empleo de *software* estadístico.

Puesto que todos los cálculos se realizan actualmente mediante programas informáticos, intentamos incluir la menor cantidad de fórmulas posible, presentando sólo aquellas que resultan indispensables para comprender la lógica de un procedimiento estadístico determinado.

Los docentes que usen este texto deberían complementarlo con actividades de práctica de administración y calificación de tests, con la lectura crítica de manuales de tests y artículos de revistas científicas relacionados con la psicometría en diferentes contextos aplicados de la psicología, así como con análisis de datos psicométricos en programas estadísticos computarizados.

Quisiéramos agradecer a todos los colegas y estudiantes que con sus observaciones críticas y comentarios a nuestros materiales impresos previos nos permitieron mejorar la claridad conceptual y expositiva de este manual. Un reconocimiento especial a Leonardo Medrano, ayudante alumno de la cátedra, quien prestó una valiosa colaboración en la elaboración de los gráficos incluidos en este texto. También resultó esencial la lectura crítica realizada por el profesor Livio Grasso, cuyas observaciones y sugerencias fueron especialmente útiles para mejorar la inteligibilidad del texto.

SILVIA TORNIMBENI
EDGARDO PÉREZ
FABIÁN OLAZ
Facultad de Psicología
Universidad Nacional de Córdoba

Primera Parte

FUNDAMENTOS DE LA MEDICIÓN EN PSICOLOGÍA

1
PROBLEMÁTICA DE LA MEDICIÓN PSICOLÓGICA

Silvia Tornimbeni - Fabián Olaz - Edgardo Pérez

1.1. La medición en psicología

Si bien muchos textos de psicometría comienzan definiendo los tests psicológicos, creemos necesario partir de un concepto previo y más general de medición. El problema de la medición posiblemente es más controversial en psicología que en otros dominios del conocimiento, debido a la complejidad del comportamiento humano y las limitaciones de los instrumentos utilizados en esa disciplina. En la actualidad aún no existe consenso en la comunidad psicológica acerca del estatus teórico de la medición.

Estas divergencias se originan en diferentes posturas filosóficas referentes al conocimiento científico y las formas óptimas de construirlo. Las distintas posiciones teóricas acerca de la naturaleza de la "medición auténtica" varían de acuerdo a las diversas concepciones acerca de la ciencia, las cuales a su vez están determinadas por diferentes enfoques sobre la naturaleza humana y la realidad.

Analizando la historia y la filosofía del concepto de medición en psicología, se pueden distinguir dos modelos fundamentales: el clásico y el representacional, cada uno con diferentes perspectivas sobre el significado general de la medición y el estatus científico de la medición psicológica.

Modelo clásico

En el intento de construcción de una metodología objetiva, algunos científicos sociales han adoptado una actitud de plena

aceptación del paradigma de las ciencias naturales. En 1940, en un seminario de expertos en medición psicológica (Campbell, 1938) se elaboró un informe en el cual se ponía en duda la posibilidad de medir atributos psicológicos, debido a que en este campo no existe isomorfismo entre las operaciones de medida y las magnitudes de la propiedad a medir (Muñiz, 1998). El concepto de isomorfismo corresponde a la equivalencia entre el orden y la distancia de los niveles de una propiedad cualquiera y del sistema numérico utilizado para medirla.

Por medición se entiende la observación de propiedades cuantitativas, tales como las frecuencias o concentraciones (Bunge y Ardila, 2002). En el modelo clásico (no debe confundirse con la teoría clásica de los tests que se desarrolla al final de este libro) se postula que, para ser mensurables, esas propiedades deben poseer las características de una variable cuantitativa. Desde esta perspectiva, propiedades tales como masa o peso pueden ser medidas, pero la medición de otras como personalidad o inteligencia, por ejemplo, es más problemática puesto que no son variables estrictamente cuantitativas.

Para que una variable sea cuantitativa debe poseer las características de distintividad, orden, aditividad y proporcionalidad (se pueden realizar juicios del tipo $A + B \geq C + D$). Sólo en el caso de que se pueda demostrar en forma empírica que una propiedad posee estas características, podríamos hablar de medición en sentido estricto.

Según Campbell (1938), la medición puede ser fundamental o derivada. Estas categorías determinan el significado de los símbolos numéricos empleados para medir. Las mediciones fundamentales no requieren otras medidas para ser expresadas (por ejemplo, las de variables como la longitud o el peso). Las variables medidas "fundamentalmente" poseen significado constitutivo y operacional por sí mismas, es decir que uno no debería "asignar" números para medir una propiedad sino "descubrir" su magnitud. Por el contrario, las mediciones derivadas son aquellas que para poder ser expresadas necesitan de otras medidas (para medir la densidad es necesario conocer previamente el volumen y la masa, por ejemplo).

Campbell afirma que la medición fundamental o directa debe ocupar un lugar central en toda disciplina que pretenda ser

científica. Esta concepción restrictiva limita la medición en psicología, puesto que la mayoría de las escalas utilizadas en esta disciplina no poseen cero absoluto, las propiedades medidas casi nunca son isomórficas con el sistema numérico y, por consiguiente, las operaciones de medición son casi siempre derivadas.

Otros autores contemporáneos, tales como Kline (2000) y Mario Bunge (1983), podrían ser incluidos en este modelo clásico de medición aunque no adhieran a una concepción tan restrictiva como la formulada por Campbell. Bunge afirma que *cuantificar* significa proyectar el conjunto de grados de una propiedad sobre un conjunto de números de modo tal que la ordenación y el espaciamiento de los números refleje el orden y el espaciamiento de los grados de la propiedad; y *medir* significa determinar efectivamente algunos de esos valores numéricos mediante el uso de una escala. Para Bunge, *la medición propiamente dicha requiere escalas con cero absoluto y unidades de medidas que pertenezcan a un sistema teóricamente fundado.*

No obstante, Bunge y Ardila (2002) reconocen que en la mayoría de los casos, en ciencia, las propiedades a medir son inaccesibles a la observación directa (las capacidades mentales o las masas atómicas, por ejemplo). Cuando la medición es indirecta debe realizarse utilizando indicadores operacionales adecuados, es decir: "propiedades observables *legalmente ligadas* a otras inobservables" (p. 83). En ese sentido, la concentración de noradrenalina en sangre sería un indicador (observable) del estrés (inobservable) o, del mismo modo, el movimiento rápido del ojo un indicador del sueño.

El problema en psicología, para Bunge y Ardila, es que muchos constructos –es decir, conceptos teóricos que no son directamente observables– y sus indicadores operacionales no han sido definidos ni explicados claramente por teorías científicas y, por consiguiente, la medición no sólo es indirecta (lo cual no sería un problema grave) sino meramente empírica y ambigua. En realidad, no existe una clasificación objetiva y fiable de la inteligencia o la personalidad; de hecho, una de las características de la psicología contemporánea es su fuerte fragmentación en "sistemas" o "escuelas" rivales. Como veremos más adelante, constructos importantes para la psicología son definidos de manera diferente por teorías "competidoras". Coincidentemente,

Kaplan y Saccuzzo (2006) expresan que los tests psicológicos no pueden ser mejores que las teorías y supuestos en los que se basan. Para Kline (2000), los tests psicológicos no son instrumentos científicos como los utilizados en las ciencias naturales (puesto que carecen de cero absoluto, unidades de medición significativas y no miden variables cuantitativas), aunque poseen un indiscutible valor pragmático en la psicología aplicada (ocupacional o educacional, por ejemplo); por consiguiente, no deberían ser abandonados hasta que la psicología disponga de teorías biológicas y cognitivas válidas que le permitan elaborar herramientas de medición superiores a los tests.

Modelo representacional

En el modelo representacional, los números utilizados en la medición no representan propiamente cantidades sino relaciones (Mitchell, 1990; Stevens, 1949). Este enfoque distingue entre un sistema relacional empírico (X), un sistema relacional numérico (R) y una aplicación de X en R. El sistema relacional empírico hace referencia al conjunto de indicadores de un constructo y las relaciones entre los mismos y el sistema relacional numérico; al conjunto de números y sus relaciones, los que pueden ser usados para representar las relaciones observadas entre los objetos o propiedades (Aftanas, 1988). En el contexto de este modelo, medir significa utilizar el sistema numérico para representar relaciones empíricas (asignar números) aunque no exista isomorfismo entre ambos sistemas.

Para comprender la afirmación precedente, consideremos diferentes clases de relaciones empíricas. El primer tipo es la relación de equivalencia, esto es, los objetos son equivalentes en una propiedad determinada, por lo cual forman parte de una misma categoría, y difieren en esta propiedad de los miembros de otras categorías. Por ejemplo, consideremos el caso de una clasificación por zona de residencia (urbano-rural), en donde asignamos un 1 a la categoría urbano y un 2 a la categoría rural. La escala de medición utilizada para representar relaciones de equivalencia se denomina *nominal*, y como se aprecia en el ejemplo, la operación básica es la clasificación. Las categorías

deben ser exhaustivas (abarcar todos los objetos que incluyen) y mutuamente excluyentes (un objeto no puede estar en más de una categoría). En esta escala los números asignados a cada categoría no representan más que una etiqueta, de forma tal que podríamos utilizar letras o cualquier otro símbolo (en lugar de números) para diferenciar un grupo de otro. Las únicas operaciones numéricas permitidas en este nivel de medición son el modo (para representar la tendencia central), los coeficientes de contingencia (para las relaciones entre variables) y las distribuciones de frecuencia.

El segundo tipo de relación es la de orden, vale decir que los objetos incluidos en una categoría no solamente difieren de los de otra sino que además pueden ser ordenados. Este tipo de escalamiento se denomina *ordinal,* y un ejemplo sería el nivel educativo (primario, secundario, terciario) o el estatus socioeconómico (bajo, medio bajo, medio, medio alto, alto). En cuanto a las propiedades formales, la escala ordinal incluye tanto la relación de equivalencia como la relación de orden (más grande que, o mayor que). En esta escala no existen intervalos iguales y, por consiguiente, no puede asegurarse que la distancia entre dos puntos de la escala (2 y 4, por ejemplo) sea equivalente a la existente entre otros dos (5 y 7, por ejemplo). Esto implica que operaciones como la suma y la resta no son admisibles en este nivel de medición. Las estadísticas que se admiten son la mediana y la correlación de rangos.

Un tercer nivel de medición es aquel en el cual se puede asumir la existencia de intervalos iguales en la escala de medición. Así, por ejemplo, en los primeros experimentos llevados a cabo en el campo de la psicofísica se solicitaba a un individuo que estimara si la diferencia en magnitud entre un par de estímulos era tan grande como la diferencia en magnitud entre otros dos estímulos. La escala numérica que permite representar este tipo de relación se denomina *intervalar*. En este nivel de medición tenemos categorías diferentes (como en la escala nominal), orden (como en la escala ordinal) y distancias numéricas que se corresponden con distancias empíricas equivalentes en las variables que se desea medir, aunque el origen de la escala es arbitrario (Cortada de Kohan, 1999). En una escala de intervalo, la distancia entre 2 y 4 es la misma que entre 21 y 23. La suma

y la resta son operaciones legítimas pero no así la multiplicación y división y, por consiguiente, 60 no representa el doble de 30 ni la mitad de 120 en este nivel de medición. Un ejemplo típico de escala intervalar es el termómetro Fahrenheit (donde el cero es relativo y arbitrario puesto que no indica la ausencia absoluta de calor). En psicología, en general, los resultados de los tests son tratados como datos de una escala intervalar aunque originalmente provengan de escalas ordinales. En efecto, como veremos más adelante si las puntuaciones de un test se distribuyen normalmente, la conversión de las puntuaciones originales a puntuaciones *z* resulta en unidades que pueden considerarse cuantitativamente iguales (Kerlinger y Lee, 2002). Estadísticas paramétricas como la desviación estándar, la media y el coeficiente de correlación lineal son admisibles en este nivel de medición.

Un último tipo de relación es aquel en el cual existe un cero absoluto con significado empírico, es decir que el cero en la escala de medición representa la ausencia absoluta de una propiedad. Esta escala se denomina "de razón" o "proporcional" y permite realizar todas las operaciones matemáticas, incluyendo la multiplicación y la división. Los números de una escala de razón indican las cantidades reales de la propiedad medida, y la longitud o el peso son variables que se miden utilizando escalas de este tipo. La escala de razón tiene todas las características de una escala de intervalo, pero además posee un cero absoluto o natural en su origen, por lo cual, un cambio en la unidad de medida no altera los juicios acerca de los valores absolutos de los atributos. En psicología, por ejemplo, el empleo de una escala de razón permitiría expresar que un individuo con una puntuación de 8 en un test X posee el doble de la propiedad P que otro individuo que obtuvo una puntuación de 4 en ese test. Sin embargo, este tipo de afirmaciones resultan inadecuadas para la mayoría de los tests psicológicos puesto que los datos con los que trabajan los científicos sociales no son ni siquiera aproximados a los requeridos para el uso de una escala de razón (Kerlinger y Lee, 2002).

Como puede deducirse de lo anterior, los números utilizados para representar un tipo de relación (equivalencia, por ejemplo) no pueden ser tratados estadísticamente de la misma forma que los utilizados para representar otro tipo de relación (orden, por

ejemplo). Por este motivo, el coeficiente de correlación (uno de los datos estadísticos fundamentales de la psicometría) y todos los métodos relacionados (análisis factorial, análisis de regresión múltiple) sólo pueden utilizarse en escalas que alcancen (mínimamente) un nivel intervalar de medición.

> El modelo representacional admite diferentes "niveles de medición" que dependen del tipo de escala (nominal, ordinal, intervalar, proporcional) empleada para medir una propiedad. Ésta es una diferencia esencial con respecto a los autores que defienden una concepción "clásica" de medición quienes postulan que la medición, auténtica tiene lugar sólo cuando se miden variables cuantitativas utilizando una escala proporcional o de razón.

Podría concluirse que el modelo clásico representa un estándar óptimo pero difícilmente alcanzable en la actualidad, y el modelo representacional una solución de compromiso más factible en el estado actual de la psicometría. Los tests psicológicos representan un avance considerable en objetividad, confiabilidad y capacidad predictiva con relación a otros métodos de evaluación (entrevista clínica, por ejemplo) pero requieren teorías válidas y explicativas (no meramente descriptivas) de los constructos e indicadores que pretenden medir para constituirse en instrumentos plenamente científicos.

1.2. Psicometría y tests psicológicos

Por todo lo expresado anteriormente se comprenderá que una de las áreas fundamentales de la psicología es la *psicometría,* que se ocupa de los procedimientos de medición del comportamiento humano, incluyendo a los denominados tests psicológicos. Para Muñiz (2001), la teoría de los tests (que veremos en el último capítulo) es sólo uno de los campos de la psicometría, que además comprende la teoría de la medición o fundamentación teórica de las operaciones de medida (abordada sintéticamente en el apartado anterior) y la estadística aplicada a la construcción y análisis psicométrico de los instrumentos de medición.

Los tests psicológicos se construyen, en general, para medir constructos que no pueden observarse directamente. Nunnally y Bernstein (1995) afirmaron que nunca se miden las personas sino algunos de sus atributos, es decir, características particulares de los individuos. En psicología nadie se propone "medir" un niño, sino su inteligencia, estabilidad emocional o autoestima, por ejemplo. Por otro lado, las operaciones de medición en psicología son casi siempre indirectas, vale decir, suponen la determinación de los indicadores del fenómeno a medir.

Como argumentó Martínez Arias (1995), el estatus actual de la psicología genera una serie de dificultades para el desarrollo de instrumentos científicos de medición, a saber:

a) Un mismo constructo psicológico puede ser definido de manera diferente, por lo cual distintos procedimientos de medida pueden conducir a inferencias disímiles en relación a aquél.
b) Es difícil determinar las características de una muestra de elementos (ítems) de un test para que sea representativa, en cuanto a extensión y variedad de contenidos, del dominio o constructo que se quiere medir.
c) Como consecuencia de lo expresado en los puntos anteriores siempre existen errores en las medidas.
d) Las escalas de medición usadas en psicología carecen, casi siempre, de cero absoluto y de unidades de medidas constantes.

Aun con estas deficiencias, el nivel de precisión alcanzado por la medición en psicología permite exhibir algunas ventajas respecto a la observación natural o no formal del comportamiento, entre ellas:

- Una de las principales es la objetividad, que implica que una afirmación fáctica es posible de verificar por otros científicos en forma independiente.
- La posibilidad de medición de las variables facilita el desarrollo de investigaciones. Según Nunnally (1991), los avances en las ciencias en general, y en la psicología en particular, se relacionan con los adelantos en los métodos de medición, aunque lo opuesto también es una realidad constatable en la historia de la ciencia.
- Los índices numéricos utilizados por los tests permiten comunicar los resultados de una evaluación con mayor precisión. De este modo, los tests proporcionan discriminaciones más sutiles que la clasificación intuitiva que un maestro podría hacer de sus estudiantes, incluyéndolos en categorías poco discriminativas como "brillante", "promedio" o "debajo del promedio", por ejemplo.
- El desarrollo de tests es un proceso complejo, pero el resultado final es un procedimiento estandarizado más sencillo y breve que la observación. Pensemos, en relación con esta última aseveración, en el tiempo requerido para administrar y puntuar un test en comparación con el tiempo que demandaría la observación del desempeño o comportamiento de una persona en su ambiente natural (escuela o trabajo, por ejemplo).

La delimitación del concepto de *tests psicológicos* no es sencilla, y a lo largo de la historia de la psicología ha suscitado innumerables polémicas. El término inglés *test* (prueba, examen) proviene del vocablo latino *testa-testis,* que denominaba una balanza utilizada en la antigüedad para pesar vasijas de oro (Cortada de Kohan, 1999).

De acuerdo con Anastasi y Urbina (1998), un test es un instrumento de medición del comportamiento de un individuo, a partir del cual pueden inferirse otros comportamientos relevantes. En 1999 la American Psychological Association (en adelante APA) definió a los tests como *un procedimiento por medio del cual una muestra de comportamiento de un dominio especificado es obtenida y posteriormente puntuada, empleando un proceso estandarizado*. Esta definición comprende no sólo a los tests de ejecución máxima, donde las respuestas son evaluadas por su corrección y calidad sino también a los de comportamiento típico (inventarios de personalidad, por ejemplo) siempre que respeten el postulado anterior.

El concepto de "evaluación" es más comprensivo que el de test y se refiere al proceso que permite integrar la información obtenida por medio de tests con la proveniente de otras fuentes,

tales como la información relacionada con la historia clínica, familiar, ocupacional o educacional de una persona.

1.3. Reseña histórica

El hecho de que las personas difieran en su comportamiento y que esas diferencias puedan medirse se ha reconocido desde los albores de la civilización. Platón y Aristóteles escribieron sobre las diferencias individuales hace más de 2000 años, y los chinos, desde la dinastía Chang (1115 a.C.), ya tenían un programa de pruebas para el ingreso de los funcionarios públicos que evaluaba destrezas importantes para la época, tales como arquería, equitación, música, escritura y matemática (Cohen y Swerdlik, 2000).

No obstante, en su acepción actual, el empleo de los tests psicológicos se inició en Europa a fines del siglo XIX. Durante la Edad Media la preocupación por la individualidad era prácticamente inexistente, permitiéndose poca libertad para la expresión y el desarrollo de la personalidad (Aiken, 2003). Es en el Renacimiento y la Ilustración cuando resurge el interés por el aprendizaje y la creatividad.

Sin embargo, recién a finales del siglo XIX se inicia el estudio científico de las diferencias individuales en lo que respecta a habilidades y rasgos de personalidad. Los tests se desarrollaron dentro del contexto de la formulación de la teoría de la evolución de las especies y las fases tardías de la Revolución Industrial, en el marco de una creciente preocupación por el aumento de la población, la mano de obra desocupada y la paulatina democratización de las escuelas.

En este contexto, surge un llamativo interés por las diferencias individuales, especialmente las de carácter hereditario, así como también por la adaptabilidad diferencial de los seres humanos a las exigencias de un entorno cambiante. Esta filosofía, denominada "darwinismo social", centraba su interés en las diferencias de naturaleza hereditaria y la adaptabilidad de los seres humanos a las exigencias de la sociedad industrial (Sternberg, 1987).

Las diferencias observadas por el astrónomo Friedrich Bessel a comienzos del siglo XIX en los registros del paso de las estrellas a través de una línea del campo visual del telescopio, realizados

por distintas personas, se convirtió en la primera evidencia de que algunas capacidades humanas podían cuantificarse. El matemático belga Adolphe Quetelet fue el primero en plantear que la teoría estadística de la probabilidad podía aplicarse a la medición del comportamiento humano (Herrera Rojas, 1998).

A estas contribuciones se sumaron las de los primeros estudiosos de la psicofísica, tales como Gustave Fechner y Ernst Weber, y los fundadores de la psicología experimental, destacándose la figura de Wilheim Wundt con su fuerte interés por medir la magnitud de propiedades psicológicas elementales a fin de formular leyes científicas. Estos autores también pusieron de manifiesto la necesidad de controlar las condiciones de prueba y tipificar los procedimientos.

No obstante, el interés de los investigadores pioneros de la medición psicológica se orientó principalmente la formulación de leyes generales que permitiesen predecir el comportamiento, y no tanto hacia la explicación de las diferencias individuales.

Es de particular relevancia la figura de Sir Francis Galton (1822-1911), primo del célebre Charles Darwin, quien a partir de sus estudios sobre la heredabilidad de la inteligencia, fue el principal responsable del inicio del movimiento psicométrico y del interés por la medición de las diferencias individuales. Este investigador inglés, interesado por el estudio de la herencia, creó un laboratorio antropométrico en Kensington, Inglaterra, donde cualquier persona podía evaluar su estatura, peso corporal, fuerza muscular, agudeza visual y otra serie de características sensoriales y motoras. Galton construyó varios tests de discriminación sensorial con la convicción de que éstos le permitían medir la inteligencia, y fue el primer investigador en adaptar algunas técnicas estadísticas para el análisis de los resultados de los tests, constituyéndose en el precursor del uso de procedimientos de análisis cuantitativos en investigación con humanos (Herrera Rojas, 1998). Con sus estudios sobre gemelos fue también uno de los fundadores de la genética del comportamiento, uno de los campos más influyentes en la psicología contemporánea (Loelhin, 1992).

En sintonía con las ideas de Galton, James Catell construyó diferentes tests de tiempos de reacción y otras funciones mentales simples. A este autor se le debe, además, la rápida difusión

de los tests en los Estados Unidos y los primeros intentos por validarlos en relación con criterios externos, es decir, comprobar si efectivamente predecían comportamientos reales diferentes de la situación de evaluación, tales como el éxito académico de los estudiantes universitarios. Sin embargo, su aporte más significativo es el de haber introducido en la literatura psicológica el término *test mental* (Muñiz, 2001).

En el año 1895, el psicólogo francés Alfred Binet publicó un artículo en el cual criticaba los tests existentes en ese momento, considerando que medían funciones muy elementales y que poseían escasa capacidad predictiva en relación con criterios externos relevantes, tales como el rendimiento académico. Binet propuso crear tests de medición de funciones mentales más complejas, tales como juicio, memoria y razonamiento. Por su parte, Wissler (1901) demostró a comienzos del siglo XX que los tests sensoriales o de reacciones mentales simples no predecían en forma adecuada el rendimiento académico de los estudiantes. Todo esto propició la creación de medidas psicológicas más semejantes a las actividades de la vida cotidiana.

En este contexto se creó la primera escala de inteligencia, que integró las experiencias anteriores e introdujo ítems relacionados con juicio, comprensión y razonamiento. Binet y Simon, a pedido del gobierno francés, utilizaron por primera vez en 1905 una escala para identificar, entre los niños que ingresaban a primer grado, aquellos que padecían debilidad mental. Esta escala consistía en 30 problemas de dificultad creciente (comprensión verbal y capacidad de razonar con materiales no verbales) y representa el desempeño típico de los niños a una edad determinada.

En 1908 estos autores desarrollaron la noción de edad mental y también una escala más refinada que se constituyó en el prototipo de los tests individuales de inteligencia. En esta escala revisada se aumentó el número de ítems y los mismos fueron agrupados sobre la base del rendimiento de una muestra grande de niños normales con edades de entre 3 y 13 años. De este modo, en el nivel (edad mental) de 3 años se agruparon todos los ítems que resolvía el 80% de los niños normales de esa edad y así sucesivamente hasta los 13 años (Binet y Simon, 1916; Anastasi y Urbina, 1998).

En la revisión de la escala Binet-Simon, realizada por Terman y conocida como Stanford-Binet, aparece la noción de Cociente Intelectual (CI). La propuesta de Terman del CI como unidad de medida de la inteligencia, con todas sus limitaciones (entre las que se destaca el hecho de que los cocientes intelectuales no serían comparables entre edades, debido a diferencias en la variabilidad de la ejecución del test), tiene una gran importancia en la psicometría, al punto tal que el CI se convirtió casi en un mito.

En esta época también fueron muy importantes los descubrimientos de un grupo de investigadores que perfeccionaron diferentes índices y modelos de análisis estadísticos, particularmente en la medición de la inteligencia. Se destacaron los trabajos de Karl Pearson (1857-1936), discípulo de Galton, quien desarrolló el coeficiente de correlación que lleva su nombre ("producto momento de Pearson"), sentando las bases para el análisis estadístico que se realiza actualmente en psicología.

Por otra parte, Charles Spearman (1927) inició una serie de investigaciones sobre las funciones cognitivas que lo llevaron al desarrollo del análisis factorial. Apoyándose en la observación de correlaciones entre tests, Spearman plantea su famosa teoría de dos factores. Según esta teoría, las puntuaciones de los tests pueden explicarse a través de dos factores: uno general, conocido como el factor *g*, que es común a todas las variables medidas, y uno específico, *s*, que sería exclusivo de cada una de esas variables. Pocos acontecimientos en la historia de los tests mentales han tenido una importancia tan grande como la formulación de la teoría de los dos factores de la inteligencia. Sobre ese fundamento se han construido numerosos tests, no sólo de inteligencia sino también de personalidad, intereses y otros constructos psicológicos. Spearman concibió también la teoría de la confiabilidad de los tests y, junto a Thorndike, el modelo estadístico de puntuaciones conocido luego como Teoría Clásica de los Tests (Martínez Arias, 1995).

La Primera Guerra Mundial generó grandes problemas para la selección y adiestramiento de millones de combatientes. El programa de selección masiva en el que se involucraron los psicólogos más capaces de la época significó una prueba de la madurez de la teoría y la técnica psicométrica. Se elaboraron los

primeros tests colectivos de inteligencia para la clasificación de grandes masas de reclutas, los célebres tests Alfa y Beta del ejército norteamericano. El test Army Alfa, elaborado por psicólogos militares dirigidos por Yerkes (1921), estaba constituido por ocho subtests que medían aspectos tales como razonamiento práctico, analogías y razonamiento matemático. El Army Beta era una versión no verbal del anterior, utilizada para la evaluación de combatientes con capacidades lingüísticas limitadas o que no eran angloparlantes.

Woodworth (en Anastasi y Urbina, 1998) desarrolló su Personal Data Sheet, un autoinforme con preguntas sobre sintomatología mental, tales como ¿usted toma whisky todos los días? (*en mi caso debo reconocer que sí*). La finalidad de este instrumento era detectar soldados con trastornos psicológicos y que no fueran aptos para el servicio militar durante la Primera Guerra Mundial. Este inventario se convirtió en modelo para los inventarios de personalidad posteriores, más sofisticados, que revisaremos más adelante. La amplia difusión de los tests colectivos durante la primera conflagración mundial fue observada con interés por los educadores, dada la practicidad de estos instrumentos. Como consecuencia de todas estas innovaciones se produjo una actividad creciente de construcción de pruebas y se desarrollaron las nociones iniciales de estandarización y validación de los tests mentales.

Es importante destacar en esta época la obra de Rorschach (1921), el psiquiatra suizo que publicó una técnica de psicodiagnóstico basada en una serie de láminas con manchas de tinta, recomendando su uso como herramienta de investigación. El test de Rorschach configuró una nueva tendencia en la evaluación psicológica vinculada con modelos teóricos psicodinámicos.

El año 1935 ha sido calificado como "bisagra" entre el período "histórico" y "moderno" dentro de la psicometría (Sternberg, 1987). Ese año se fundó la Sociedad Psicométrica por un grupo de investigadores agrupados en torno a la figura de L. Thurstone, investigador de la Universidad de Chicago. También en este año surge la primera publicación especializada en la medición psicológica, *Psychometrika*, que continúa vigente en la actualidad.

La mayoría de los tests publicados hasta ese momento se basaban en la concepción de la inteligencia como rasgo unitario.

Los investigadores nucleados en la Sociedad Psicométrica desarrollaron una innovación fundamental: el análisis factorial moderno, un método que demostraba con claridad que la inteligencia es algo más que una capacidad unitaria. El psicólogo estadounidense Thurstone realizó una serie de aportes a la lógica y los fundamentos matemáticos del análisis factorial, logros que facilitaron la medición de aptitudes más específicas, que contribuyen al desempeño cognitivo más allá de la influencia de la inteligencia general o *g*. El *test* de Aptitudes Mentales Primarias de Thurstone (1935) fue un modelo para las baterías de tests multifactoriales posteriores, inaugurando una nueva manera de concebir y medir la inteligencia**.**

Como hemos dicho, con el empleo del análisis factorial se construyeron numerosos tests, no sólo de inteligencia, sino también de personalidad, intereses y otros atributos psicológicos. Teorías contemporáneas tales como la de la inteligencia fluida (Gf) y cristalizada (Gc) de Cattell (1967), la teoría de los cinco factores de la personalidad (Norman, 1963; Costa y Mc Crae, 1999; Goldberg, 1999) y otros modelos semejantes (Caroll, 1993) constituyen un refinamiento de los postulados precursores de Spearman y Thurstone.

La Segunda Guerra Mundial, con sus necesidades de incorporación de millones de reclutas, también estimuló la construcción de tests de aptitudes específicas, que fueron muy útiles para seleccionar pilotos, bombarderos, operadores de radio y otras funciones militares especializadas. Por esa época, Guilford (1967) construyó para la fuerza aérea una batería de tests que medían diferentes factores de la estructura de la inteligencia humana. El Test de Aptitudes Diferenciales (Bennet, Seashore y Wesman, 2000), entre otros similares, son herederos de esos descubrimientos.

La década de 1950 es considerada como una fase "madura" de la teoría de los tests, puesto que aparecieron textos que con el tiempo serían clásicos y dejarían establecidos los fundamentos teóricos de la psicometría. Surge en ese momento histórico una corriente de revisión y análisis de la fundamentación científica de las pruebas. Los trabajos realizados en este período versan en su gran mayoría sobre teoría de la medición, los principios y fundamentos de la medición en psicología, los proble-

mas de validez y confiabilidad y, en síntesis, la construcción de una teoría psicométrica. Así, pueden mencionarse *Theory of Mental Tests* (Gulliksen, 1950) y las normas técnicas iniciales de la APA, entre otras obras valiosas (Hogan, 2004).

En la década de 1960 se comenzó a criticar esta concepción clásica de la teoría de los tests, al tiempo que aparecían teorías alternativas. Hay dos modelos originados en esa época que prevalecen en la literatura psicométrica actual: el de maestría de dominio y el de rasgo latente. Dentro del primero se ubican los denominados tests con referencia a criterio, término introducido por Glaser (1963), que miden un dominio de conocimiento claramente delimitado. Estos tests están íntimamente ligados al campo educativo. Por otra parte, la teoría de rasgo latente (Rasch, 1963) derivó en la Teoría de Respuesta al Ítem (Lord, 1980), uno de los paradigmas relevantes de la psicometría contemporánea. Ambos enfoques serán revisados más adelante.

En los últimos años del siglo XX se produjo un acercamiento entre la psicometría y la psicología cognitiva, y se elaboraron modelos psicométricos denominados "modelos componenciales" que incorporan los diferentes componentes de los procesos cognitivos en la resolución de un problema (Van der Linden y Hambleton, 1997; Prieto y Delgado, 1999). Estos modelos también se conocen como "evaluación inteligente"; en ellos se presentan tareas que son comunes en la vida real. Un modelo componencial requiere: a) un análisis de las operaciones mentales (componentes cognitivos) que intervienen en la resolución de los ítems y b) un modelo matemático que estime la probabilidad de responder correctamente el ítem teniendo en cuenta sus propiedades psicométricas y el nivel de conocimiento del sujeto.

El uso de las computadoras en psicometría tuvo un notable incremento desde la década de 1980, aplicándose en casi todas las instancias de la evaluación psicológica. Debido a su consistencia, la computadora lleva al extremo la estandarización y objetividad de un test. Algunas de las aplicaciones más interesantes de la informática en los tests psicológicos son:

a) *Bancos de ítems*: La forma de presentación más habitual de un test es un cuadernillo impreso con los ítems a resolver. Sin embargo, un test bien ajustado para determinados propósitos puede ser rápidamente creado a partir de un banco de ítems. Una escuela puede solicitar a un editor algunos tests adecuados a los contenidos de su currícula. En la actualidad es posible elaborar un banco de ítems y, a partir del mismo, construir tests con una computadora. Otra posibilidad es generar un número ilimitado de formas equivalentes de un test, seleccionándolas con el mismo criterio de un banco de ítems. Un procedimiento denominado GAI (generación automática de ítem) permite generar ítems mediante determinados algoritmos, que requieren programas específicos como el Rasch Item Calibration Program (RASCAL, 1989) u otros semejantes. La Teoría de Respuesta al Ítem, revisada al final de este texto, constituye el marco conceptual y metodológico para el desarrollo de bancos de ítems de tests.
b) *Administración y puntuación asistida por computadora*: Las computadoras son adecuadas para administrar y puntuar los tests. Al aplicar una prueba en formato computarizado se obtienen mediciones precisas e instantáneas, no hay errores en la corrección y se consiguen informes legibles con posibilidad de transmisión y multicopias impresas. Además la interacción con las computadoras fascina a las nuevas generaciones y se espera que esto vaya en aumento creciente. Un inconveniente es que la informatización puede ocasionar una pérdida de la riqueza de las observaciones no formales que realizan los administradores expertos durante la aplicación de un test individual, algo que puede atenuarse si el administrador acompaña el proceso de respuesta del individuo al test computarizado (Cronbach, 1998).
c) *Software de simulación*: Las nuevas tecnologías incrementan notablemente la variedad de los estímulos incluidos en los tests. Los simuladores de vuelo, por ejemplo, representan de modo realista el instrumental que deben manipular los pilotos y proporcionan continua retroalimentación de los resultados de sus operaciones. Aunque fueron

diseñados para entrenamiento, estos dispositivos pueden también ser empleados para la evaluación del progreso en el aprendizaje en cualquier dominio.

La mayoría de los tests empleados en las diversas áreas de la psicología disponen de versiones computarizadas. El directorio de *software* psicológico de la APA (1999) describe regularmente centenares de programas para administrar y/o interpretar tests por computadoras. Estas tecnologías permiten economizar el proceso de puntuación y elaboración de perfiles y, al mismo tiempo, mejoran la precisión y objetividad de los tests convencionales (de lápiz y papel) al eliminar los errores que se cometen durante la puntuación manual de los mismos.

El alcance de la informática en psicometría no se limita al empleo de tests asistidos por computadora. En las últimas décadas se ha diseñado una amplia variedad de programas estadísticos que incluyen rutinas y menúes adecuados para resolver problemas de investigación en este dominio (estudios correlacionales, análisis factorial, entre otros). En el apéndice de este libro presentamos ejemplos de algunos análisis psicométricos realizados con *software* estadístico moderno.

La búsqueda de información relacionada con los tests también se ve sumamente facilitada por los recursos disponibles en Internet, donde se encuentran bases documentales de gran utilidad como la que ofrece el sitio web de la American Psychological Association (www.apa.org), así como revistas y portales científicos *on line* (www.sciencedirect.com, por ejemplo), y editoriales abocadas exclusivamente a la publicación de tests, tales como TEA en España (www.teaediciones.com).

En la Facultad de Psicología de la Universidad Nacional de Córdoba existe una revista electrónica especializada, *Evaluar*, que periódicamente publica trabajos teóricos y empíricos relacionados con la medición psicológica y educativa (www.revistaevaluar.com.ar).

En síntesis, la psicometría moderna evidencia tres características fundamentales: a) la importancia de la teoría de respuesta al ítem en la construcción de tests, coexistiendo con la teoría clásica de los tests (y en algunos casos reemplazándola); b) la presencia creciente de los tests basados en computadora en

lugar de los tests de lápiz y papel, y c) el diseño de ítems más atractivos y realistas, que incorporan recursos audiovisuales y de simulación computarizada, hecho que promete una nueva generación de tests de mayor validez (Kaplan y Saccuzzo, 2006; Moreno, Martínez y Muñiz, 2004).

2
CLASIFICACIÓN DE LOS TESTS

Edgardo Pérez

En la literatura psicométrica encontramos diversas taxonomías que utilizan criterios disímiles para clasificar los tests psicológicos. Así, por ejemplo, éstos suelen agruparse en: a) *individuales o grupales*, según se administren a una persona por vez o a un grupo de individuos simultáneamente; b) de *ejecución, lápiz y papel, visuales, auditivos, o computarizados*, de acuerdo al formato y materiales de presentación de los tests, o c) basados en la *teoría clásica o de respuesta al ítem*, conforme al modelo teórico de construcción. Cronbach (1998) distinguió entre tests de *ejecución máxima y* medidas de *ejecución o respuesta típica*, según demanden el mayor rendimiento del examinado (como acontece en los tests de habilidades) en sus respuestas, o midan el comportamiento habitual sin requerir respuestas correctas (a la manera de los inventarios de personalidad, por ejemplo).

Otra clasificación interesante es la propuesta por Nunnally (1991), en función de las áreas del contenido (constructos) medido por los diferentes tests. Este tipo de taxonomía es particularmente estimulante para quien se inicia en el estudio de los tests, debido a que aporta una idea general de la diversidad de los campos de aplicación en los que pueden ser utilizados. Siguiendo este criterio, Nunnally (1991) discriminó tres categorías de tests: de *habilidades,* de *rasgos de personalidad,* y de *preferencias (*intereses, valores y actitudes). No obstante, esta clasificación es problemática puesto que las diferencias conceptuales entre rasgos de personalidad y preferencias no son claras ni aceptadas unánimemente, con constructos (personalidad-intereses, intereses-actitudes, intereses-valores, por ejemplo) que

se solapan en grado considerable (Anastasi y Urbina, 1998; Holland, 1997). Por otro lado, no existen diferencias formales entre las escalas que miden actitudes, rasgos de personalidad o intereses vocacionales. En efecto, casi todas estas escalas han adoptado un formato *likert* de respuesta (*Acuerdo-Desacuerdo* o *Muy seguro-Nada seguro*, por ejemplo) que solo varía en el número de alternativas contempladas (tres, cinco, siete o diez, entre las más comunes).

Por estas razones, consideramos que la clasificación planteada por Cronbach (1998) continúa siendo la más adecuada puesto que se refiere a diferencias esenciales entre los tests incluidos en sus dos categorías (ejecución máxima y respuesta típica). En este capítulo realizaremos algunos agregados a esa clasificación clásica. En efecto, incluimos en nuestra revisión la medición de las creencias de autoeficacia (Bandura, 1987; 1997) y las habilidades sociales, constructos que hoy no pueden ignorarse dada su importancia conceptual y empírica. Además, comentaremos ciertas teorías relevantes en relación con cada constructo y mencionaremos tests psicológicos desarrollados internacionalmente y en nuestro ámbito en estas dos últimas décadas.

2.1. Tests de ejecución máxima: inteligencia, aptitudes y habilidades

La característica principal de los tests de ejecución máxima es que demandan a los examinados que respondan de la forma más eficiente que puedan frente a tareas problemáticas (problemas matemáticos, por ejemplo) que deben resolver (Cronbach, 1998). En estos tests se miden diferencias individuales en el nivel de ejecución máximo ante distintas tareas, cuando se intenta realizarlas (Nunnally, 1991). Esto significa que los desempeños solamente pueden medirse cuando las personas están motivadas para realizar una tarea de la mejor manera posible.

Bajo el concepto genérico de tests de ejecución máxima se incluyen variables relacionadas, tales como las aptitudes, las habilidades y la inteligencia. Debe aclararse que la delimitación de estos conceptos es uno de los problemas más controvertidos de la psicología, al igual que el dilema subyacente de la determinación genética o cultural del comportamiento. Para Juan-Espinosa (1997), una habilidad desarrollada representa el logro en algún dominio (por ejemplo, la escritura) y la inteligencia, una condición necesaria para ese logro. De acuerdo con este autor, la inteligencia general y las aptitudes específicas (verbal, espacial o matemática, por ejemplo) dependen de características ligadas a la constitución cerebral y de disposiciones genéticas de las personas, y son más resistentes al entrenamiento que las variables medidas por los tests de logro o habilidades desarrolladas. No obstante, en algunos tests de inteligencia o aptitudes se incluyen ítems que parecen medir habilidades desarrolladas más que aptitudes.

La postulación de un factor cognitivo general (*g*) que permite resolver problemas novedosos de cualquier naturaleza se opone a la concepción de aptitudes relativamente independientes, también tradicional en la psicología. La existencia de un factor general de inteligencia es apoyada por investigaciones psicométricas y de la genética del comportamiento (Plomin, DeFries, McClearn y McGuffin, 2002), pero esto no implica negar la existencia de aptitudes más específicas. En general se asocia este factor *g* a la velocidad de procesamiento cognitivo cuyas bases biológicas no están aún bien determinadas, aunque se ha encontrado alguna evidencia preliminar en relación con la velocidad de conducción nerviosa y el número de neuronas corticales, entre otros indicadores psicobiológicos. Se ha definido la inteligencia general como flexibilidad comportamental y mental para encontrar soluciones novedosas a problemas. Claramente, la inteligencia no es exclusiva de la especie humana aunque el hombre sea el mamífero más inteligente (Roth y Dicke, 2005).

Un test que se considera un indicador adecuado de *g* es el de Matrices Progresivas de Raven (1993). Se trata de una prueba no verbal, cuyos ítems muestran un patrón de relaciones (cruces y círculos, por ejemplo) incompleto, donde los examinados deben responder seleccionando la secuencia faltante que completa la serie. Si bien sus autores aseguran que este test mide "educción de relaciones", un concepto estrechamente relacionado con la inteligencia general, diversos análisis factoriales han cuestionado esta estructura interna unitaria del test. Se ha afirmado que el Raven, en realidad, mide tres factores cognitivos (percepción,

razonamiento analógico y capacidad espacial) y que debería complementarse con una medida del razonamiento verbal para ofrecer un perfil más completo de la inteligencia en relación con las teorías actuales (Hogan, 2004).

Howard Gardner (1994, 1999) efectuó una crítica radical al modelo de inteligencia general con su Teoría de las Inteligencias Múltiples (*Multiple Intelligences*, MI). Para Gardner, los tests miden preferentemente aptitudes relacionadas con los requerimientos académicos de la cultura occidental y por eso sólo identifican dos o tres dimensiones (lingüística, espacial y lógico-matemática) de la inteligencia. Su teoría, basada primordialmente en criterios neuropsicológicos, propone ocho potenciales biopsicológicos de procesamiento de información ("inteligencias") que permiten resolver problemas o crear productos valorados por una cultura. Estas inteligencias, según Gardner (1999), son: Lingüística, Lógico-Matemática, Espacial, Cinestésico-Corporal, Musical, Interpersonal, Intrapersonal y Naturalista. La teoría MI, de fuerte atractivo entre los educadores, constituye una fuente riquísima de hipótesis que no poseen una corroboración empírica exhaustiva ni técnicas objetivas de medición de sus constructos e indicadores (Hood y Johnson, 2002). En efecto, los tests construidos para medir aspectos relacionados con las inteligencias múltiples, tales como el Multiple Intelligence Developmental Assessment –MIDAS– (Shearer, 1999) o el Inventario de Autoeficiencia para Inteligencias Múltiples –IAMI– (Pérez, 2001), evalúan habilidades autopercibidas o autoeficacia (concepto que trataremos más adelante) para actividades relacionadas con las ocho inteligencias.

También existen desarrollos teóricos contemporáneos que representan una solución de compromiso entre ambas posturas, admitiendo la existencia del factor *g* pero también de aptitudes y habilidades relativamente independientes. Una de estas teorías es la de Cattell-Horn-Carroll (CHC) (Carroll, 1993; McGrew, Flanagan, Keith y Vanderwood, 1997), que propone un modelo de tres estratos: la inteligencia general en el estrato superior (*g*), un estrato medio de aproximadamente diez aptitudes cognitivas (procesamiento visual, por ejemplo) y un estrato inferior con numerosas habilidades más específicas (como las destrezas manuales). Un instrumento contemporáneo basado explícitamente en la teoría CHC es la batería Woodcock-Johnson-III (WJ-III) de aptitudes cognitivas (Woodcock, McGrew y Mather, 2001). Las aptitudes medidas por este test son: rapidez en el procesamiento, procesamiento visual, procesamiento auditivo, memoria, comprensión-conocimiento, razonamiento fluido, lectura-escritura y aptitud cuantitativa. Existen versiones de la WJ-III en varios idiomas (incluida una versión en español) y con un rango de aplicación de 2 a 90 años. Este test es de administración individual y posee buenas propiedades psicométricas de estandarización, confiabilidad y validez.

La teoría CHC representa un notable esfuerzo para lograr la conceptualización de la inteligencia. Sin embargo, aún existen desacuerdos básicos entre los defensores de esta teoría. Por ejemplo, algunos investigadores aceptan la existencia de *g* como un tercer estrato mientras que otros hablan sólo de dos estratos (aptitudes amplias y habilidades específicas). Del mismo modo, no existe consenso respecto de la cantidad de aptitudes del segundo estrato.

Recientemente (Johnson y Bouchard, en prensa) se ha propuesto otro modelo alternativo de la estructura de la inteligencia humana, el VPR (verbal-perceptual-rotación de imágenes), basado en la teoría originalmente formulada por Vernon (1964). La teoría VPR propone un factor general de inteligencia, un segundo estrato de tres aptitudes generales (verbal, perceptual y de rotación de imágenes) y un tercer estrato de ocho aptitudes más específicas relacionadas con las anteriores (verbal, académica, fluidez, numérica, memoria, espacial, velocidad perceptiva y rotación de imágenes). Este modelo se basa en evidencias psicométricas, neurocientíficas y provenientes de la genética del comportamiento. Estas últimas indican que un 70% de la variabilidad de esta estructura de la inteligencia es explicada por factores genéticos.

En síntesis, la investigación parece apoyar la existencia de un factor general de inteligencia, que no explica la variabilidad total del comportamiento inteligente, y de aptitudes cognitivas que realizan una contribución específica al comportamiento inteligente, más allá de la contribución de *g*. Las aptitudes de mayor relevancia consensuadas en las diferentes teorías son las denominadas verbal y espacial; las demás aptitudes generales y

específicas asociadas constituyen todavía un dominio altamente controversial.

La revista *Intelligence* es una de las publicaciones más autorizadas en relación con la investigación y medición de la inteligencia y allí regularmente aparecen artículos relacionados con las diferentes teorías que hemos mencionado.

Las escalas más utilizadas para la medición de la inteligencia en nuestro medio son las elaboradas por David Wechsler en 1939, con varias actualizaciones posteriores; las últimas referentes al WISC-IV (Wechsler, 2005), para niños y adolescentes, y el WAIS-III (Wechsler, 1999), para adultos. Todas las escalas de Wechsler comprenden subtests verbales y de ejecución. Los ítems de los subtests verbales plantean problemas del tipo *¿Qué significa arrogante?*, o *Menciona un planeta de nuestro sistema solar que no sea la Tierra*; los subtests no verbales consisten, por ejemplo, en ensamblar objetos a la manera de un rompecabezas.

Análisis psicométricos contemporáneos de las escalas Wechsler identificaron cuatro factores de inteligencia subyacentes (organización perceptual, memoria de trabajo, comprensión verbal y velocidad de procesamiento). En la última versión del WISC-IV las puntuaciones se interpretan en función de esos cuatro factores y no en la forma tradicional de inteligencia verbal y de ejecución. En las versiones actuales de las escalas Wechsler los ítems están ordenados según los parámetros de dificultad y discriminación de la teoría de respuesta al ítem (Hogan, 2004).

La tabla 2.1. presenta un listado de los subtests de estas escalas y su relación con los cuatro factores subyacentes a las puntuaciones.

Las escalas Wechsler son muy empleadas en psicología clínica y educacional y han sido estandarizadas cuidadosamente en los Estados Unidos y España, entre otros países, con muestras nacionales representativas y estratificadas por edad, sexo, raza, educación y ocupación.

La orientación de carrera y la selección de personal son áreas de trabajo del psicólogo donde resulta de significativa importancia la medición de aptitudes cognitivas. En estos ámbitos son muy empleadas pruebas multifactoriales como el Test de Aptitudes Diferenciales (DAT-5) (Bennet, Seashore y Wesman,

Tabla 2.1. Relaciones entre los subtests y los cuatro factores de las escalas Wechsler

Puntuaciones de Índice	Comprensión Verbal	Memoria de Trabajo	Organización Perceptual	Velocidad de Procesamiento
Subtest verbales				
Vocabulario	x			
Analogías	x			
Aritmética		x		
Retención de Dígitos		x		
Información	x			
Sucesión de Letras y Números		x		
Subtests de ejecución				
Completamiento de figuras			x	
Dígitos y Símbolos Claves				x
Diseño con Cubos			x	
Matrices			x	
Búsqueda de Símbolos				x

2000), compuesto por ocho subtests que permiten obtener puntuaciones en competencias requeridas para el éxito académico u ocupacional (aptitud verbal, numérica, espacial, abstracta, mecánica, administrativa, lenguaje y ortografía).

Se ha criticado a este tipo de tests su falta de poder predictivo diferencial –puesto que los mejores predictores resultan ser los puntajes combinados de sus subtests verbales y numéricos, algo equivalente a un test de inteligencia aunque innecesariamente más extenso–, así como la confusión conceptual de incluir aptitudes (como las administrativas) que en realidad son un compuesto de factores cognitivos y de personalidad (Kline, 2000). No obstante, a los fines de orientación o selección suministran información más específica que los tests de inteligencia general, y tal vez en esto radique su popularidad entre los orientadores. El desarrollo de pruebas de aptitudes con bases científicas sólidas (en especial tests colectivos) es uno de los grandes desafíos

del futuro para la orientación vocacional y la selección de personal, dos de las áreas más importantes de la psicología aplicada (Johnson y Bouchard, en prensa).

En nuestro país, Cortada de Kohan (1998) elaboró el Test de Aptitud Verbal Buenos Aires, que consta de 98 ítems divididos en dos mitades: sinónimos y definiciones. Todos los ítems son de opción múltiple con 4 alternativas de respuesta, de las cuales una es la correcta. El tiempo de administración es libre, pero suelen ser suficientes 25 minutos para terminar la prueba, que puede ser aplicada tanto en forma individual como colectiva. El Test Buenos Aires posee baremos para la Argentina, Ecuador, Colombia y España. Se han realizado los estudios psicométricos clásicos (confiabilidad, validez, análisis de ítems), pero además se han obtenido para todos los ítems los parámetros de dificultad y discriminación según la teoría de respuesta al ítem, algo muy novedoso en nuestro país. Puede ser utilizado con adolescentes mayores, desde los 16 años, y adultos, con al menos tres años cursados de educación secundaria. También se dispone de una versión abreviada que mantiene las propiedades de confiabilidad y validez de la forma completa y que debe administrarse con un tiempo límite de ocho minutos.

Un caso especial son los tests de rendimiento o logro. Este tipo de pruebas se utilizan en todos los niveles del sistema educativo para medir el conocimiento alcanzado en un área específica. En nuestro medio, Grasso (1969) elaboró un test de conocimiento en matemática para ingresantes a la universidad. La prueba está compuesta por 70 problemas (del tipo: Si se lanzan tres monedas, ¿cuál es la probabilidad de obtener exactamente dos caras?) que el estudiante debe resolver utilizando un formato de opción múltiple de cinco alternativas de respuesta. Las propiedades psicométricas del instrumento fueron adecuadas y un análisis de regresión múltiple permitió constatar que explicaba un 76% de la varianza del rendimiento académico de los estudiantes de primer año de la Facultad de Matemática, Astronomía y Física de la Universidad Nacional de Córdoba (promedio de calificaciones). Esta contribución específica fue muy superior a la realizada por las otras variables independientes del modelo (tests de aptitudes, intereses y valores) que sólo incrementaron en un 8% la contribución explicativa del test de conocimiento en matemática respecto al rendimiento académico. Los tests referidos a criterio (TRC) constituyen un tipo especial de tests de rendimiento (en realidad, una forma particular de interpretar los resultados de estos tests) que revisaremos en el capítulo de interpretaciones de puntuaciones de tests.

Uno de los desarrollos contemporáneos más relevantes son los tests adaptativos o a medida, basados en la teoría de respuesta al ítem. En especial los tests de rendimiento educativo han comenzado a adoptar crecientemente esta modalidad. A partir de un banco de ítems, la mayoría de los tests adaptativos operan mediante una estrategia de ramificación variable para la selección progresiva de los ítems, que requiere establecer: a) un procedimiento de inicio, a partir del cual se determina el primer ítem a presentar, b) un procedimiento para seleccionar, tras una estimación provisional del nivel del individuo en el dominio, el siguiente ítem a presentar, y c) un criterio para dar por finalizada la prueba (Olea, Ponsoda y Prieto, 1999). En comparación con los tests convencionales de longitud fija, mediante un algoritmo adaptativo se consigue una mejor adecuación entre la dificultad de los ítems y el nivel de rasgo del sujeto, y por tanto se obtiene una estimación precisa de su nivel de rasgo con la presentación de pocos ítems y en un tiempo de aplicación reducido. Además, dado que diferentes individuos reciben ítems distintos, los tests a la medida previenen que los ítems no sean conocidos antes de su aplicación. Estos beneficios resultan especialmente importantes para los responsables de programas de evaluación educativa a gran escala, donde es necesario aplicar los tests de forma continua a muestras extensas.

Otro ámbito relevante para el uso de tests de ejecución máxima es la neuropsicología, que estudia las relaciones entre el cerebro y la conducta (Kolb y Wishaw, 1986). El desarrollo de la neuropsicología ha estado determinado por la necesidad de investigar y encontrar herramientas que permitan el diagnóstico y el tratamiento de los déficit en el rendimiento cognitivo (memoria, lenguaje, atención, funciones visoespaciales, funciones ejecutivas) después de producirse una lesión cerebral. Frecuentemente, estas lesiones resultan en trastornos cognitivos que afectan el desempeño de una persona en las actividades de la vida diaria, especialmente en la esfera laboral. Por ello, luego de

una lesión cerebral es imperioso determinar la cantidad y calidad de daño cognitivo que puede haber sufrido la persona.

La evaluación neuropsicológica (EN) es la herramienta que posibilita este diagnóstico. Lezak (1995) identificó cuatro aplicaciones esenciales de la EN: evaluación propiamente dicha, cuidado del paciente y planificación del tratamiento, rehabilitación y evaluación del tratamiento, e investigación. En el texto clásico de Lezak se ha realizado la mayor recopilación y descripción de tests neuropsicológicos existentes, mencionándose más de 500 pruebas de este tipo.

Las áreas cognitivas evaluadas por los tests neuropsicológicos son de una enorme variedad. Así, podemos citar, entre otras, memoria, atención, discriminación visual, gnosias visuales, gnosias auditivas, discriminación de color, funciones ejecutivas (planeamiento, verificación), lenguaje (expresión, comprensión, denominación), praxias (constructivas, de miembros). Algunos de los tests más conocidos en este ámbito son: el Mini-Mental State Examination (Folstein, Folstein y McHugh, 1975), un test de inspección rápida (dura aproximadamente 5 minutos) del estado cognitivo general de una persona; el Test de Stroop (1935), una prueba de atención que requiere determinar el color en el que están escritos los nombres de colores que se hallan impresos en colores incongruentes con la palabra (por ejemplo, la palabra "rojo" escrita en tinta verde); el Test de Clasificación de Cartas de Wisconsin (Heaton, Chelune, Talley, Kay y Curtiss, 1991), una prueba de flexibilidad cognitiva; y la Figura Compleja, de Rey (1941), un test de memoria visual y praxias constructivas. Se ha demostrado acabadamente la importancia, utilidad y justificación de esta área de evaluación, cuyo logro más reciente es la posibilidad de identificar precozmente déficit cognitivos, tales como la demencia.

Otro dominio íntimamente relacionado con la prevención e intervención es el desarrollo infantil. Los tests de evaluación del desarrollo infantil miden las áreas motora, afectiva, cognitiva y del lenguaje, facilitando la detección precoz de posibles trastornos. La población meta de estos instrumentos es la que posee entre 0 y 5 años; incluye por lo tanto la evaluación del neonato (los primeros 30 días de la vida extrauterina); el lactante (desde los 30 días hasta los 24 meses de edad) y el pre-escolar (desde los 2 a los 5 años). Estas pruebas requieren un buen entrenamiento del evaluador en el manejo, observación de niños pequeños y también sólidos conocimientos teóricos que permitan otorgar a las conductas observadas la debida importancia en el contexto de un diagnóstico. Debe destacarse que en ningún caso los tests de desarrollo reemplazan el examen neurológico del niño, sino que lo complementan.

En general, los tests de evaluación del desarrollo poseen menos confiabilidad y validez que otros tests de ejecución máxima, debido quizá a la pobre capacidad de concentración de los niños pequeños y a la rápida maduración cognitiva que caracteriza a este período de la vida (Aiken, 2003). No obstante, estos tests son útiles para el diagnóstico precoz del retraso mental, los trastornos cerebrales orgánicos y los trastornos del aprendizaje (por ejemplo, dislexia y discalculia). Entre los principales instrumentos que se utilizan en nuestro país podemos destacar las escalas de Gesell y Amatruda (1971), construidas para diagnosticar si los niños alcanzan parámetros adecuados de desarrollo. A lo largo de un extenso programa de investigación se obtuvieron datos normativos sobre el desarrollo de las habilidades motrices, lingüísticas y sociales, así como del comportamiento adaptativo, en niños de 0 a 6 años. Las puntuaciones de estas escalas, determinadas por la presencia o ausencia de conductas específicas características a determinada edad, se expresan en términos de la edad de desarrollo.

Otro instrumento de este tipo, de gran aceptación internacional, son las Escalas Bayley del Desarrollo Infantil. Las tres escalas (motora, social y comportamental) se consideran complementarias y suministran una contribución interesante a la evaluación clínica del niño (Bayley, 1993).

2.2. Tests de comportamiento típico: motivación, actitudes y personalidad

En este tipo de tests ninguna respuesta puede ser calificada como correcta o incorrecta. Aquí se evalúa el comportamiento habitual de los individuos, recurriendo a distintas afirmaciones ante las cuales el examinado debe indicar su nivel de acuerdo o

agrado, por ejemplo. Los tests de respuesta típica comprenden las medidas de rasgos de personalidad, intereses y actitudes, así como de otros constructos afectivos y motivacionales relacionados, como las creencias de autoeficacia (Cronbach, 1998). Los tests de habilidades sociales también deben incluirse en esta categoría puesto que su formato habitual de respuesta es el de un autoinforme de respuesta típica y no el de un test de ejecución máxima.

Los tests de respuesta típica son, en su gran mayorí,a inventarios de autoinforme donde se demanda al individuo información sobre sí mismo. Esta medición introspectiva y basada exclusivamente en el lenguaje genera varias limitaciones importantes, tales como no ser aplicables a niños pequeños y el hecho de que sus respuestas pueden falsearse (de manera intencional o no). Si bien se han ideado procedimientos para atenuar (no eliminar) las respuestas negligentes, deshonestas o tendenciosas, los resultados de estos tests deben intepretarse con precaución y no deberían ser nunca el único criterio utilizado para tomar decisiones clasificatorias o diagnósticas respecto a las personas.

Escalas de autoeficacia

La teoría social cognitiva ha destacado el papel de la autoeficacia percibida entre las variables motivacionales y afectivas. Bandura (1997) define la autoeficacia como la creencia en las propias capacidades para realizar determinados cursos de acción. Para este eminente teórico, las creencias de las personas acerca de sí mismas son elementos clave para la determinación de su comportamiento, dado que son un elemento de gran influencia y desempeñan un rol importante en las elecciones efectuadas por las personas, el esfuerzo que invierten, la perseverancia para alcanzar metas y el grado de ansiedad y confianza que experimentan frente a las tareas de la vida.

La autoeficacia se relaciona fuertemente con los intereses vocacionales pero se trata de una relación asimétrica, puesto que, tal como se ha comprobado en numerosas investigaciones, las personas tienden a interesarse por aquellas actividades que se sienten capaces de realizar exitosamente (Lent, Brown y Hackett, 1994). La autoeficacia también se relaciona con las aptitudes, puesto que las personas se sienten más seguras de emprender aquellas actividades en las que han experimentado éxito. No obstante, sujetos con igual nivel de habilidad pueden experimentar diferente seguridad para emprender determinados cursos de acción, por lo cual la autoeficacia permite mejorar la predicción del rendimiento que realizaríamos si sólo nos guiáramos por el nivel de habilidad real. Esto es así porque el desarrollo de creencias de autoeficacia no sólo depende del éxito previo sino de otras fuentes, tales como el aprendizaje vicario y la persuasión social.

La teoría social-cognitiva del desarrollo de carrera (Lent, Brown y Hackett, 1994) propone un modelo explicativo de las interrelaciones entre rasgos de personalidad, intereses, habilidades y autoeficacia que contribuye a esclarecer el significado diferencial de estos constructos.

En la figura 2.1. pueden observarse las relaciones entre rasgos de personalidad (más básicos y ligados a lo genético) (A), las aptitudes (también hereditarias en gran parte y una de las fuentes de la autoeficacia al facilitar las experiencias de logro en un dominio) (B), la autoeficacia (más ligada al aprendizaje, relacionada con el constructo anterior pero también influida por experiencias de aprendizaje adicionales tales como la persuasión social y el aprendizaje vicario), y los intereses vocacionales (aprendidos en gran medida y relacionados directamente con la autoeficacia y las expectativas de resultados, e indirectamente con las experiencias de aprendizaje y la personalidad).

Existen algunos interrogantes respecto a la naturaleza de la autoeficacia. En efecto, si bien Bandura (1997) puntualizó claramente que se trata de un constructo aprendido y contextualmente-específico, algunos investigadores postulan que también existe un constructo de autoeficacia general, más semejante a los rasgos de personalidad, y otros han sugerido que la herencia influye de manera modesta en la autoeficacia además del papel innegable del aprendizaje (Kaplan y Saccuzzo, 2006).

El sitio web del Dr. Frank Pajares (www.emory.edu/EDUCATION/mfp), en la Universidad de Emory, Atlanta, constituye un tesoro informativo sobre teoría, investigación y medición de la autoeficacia. Se han construido escalas de autoeficacia para el

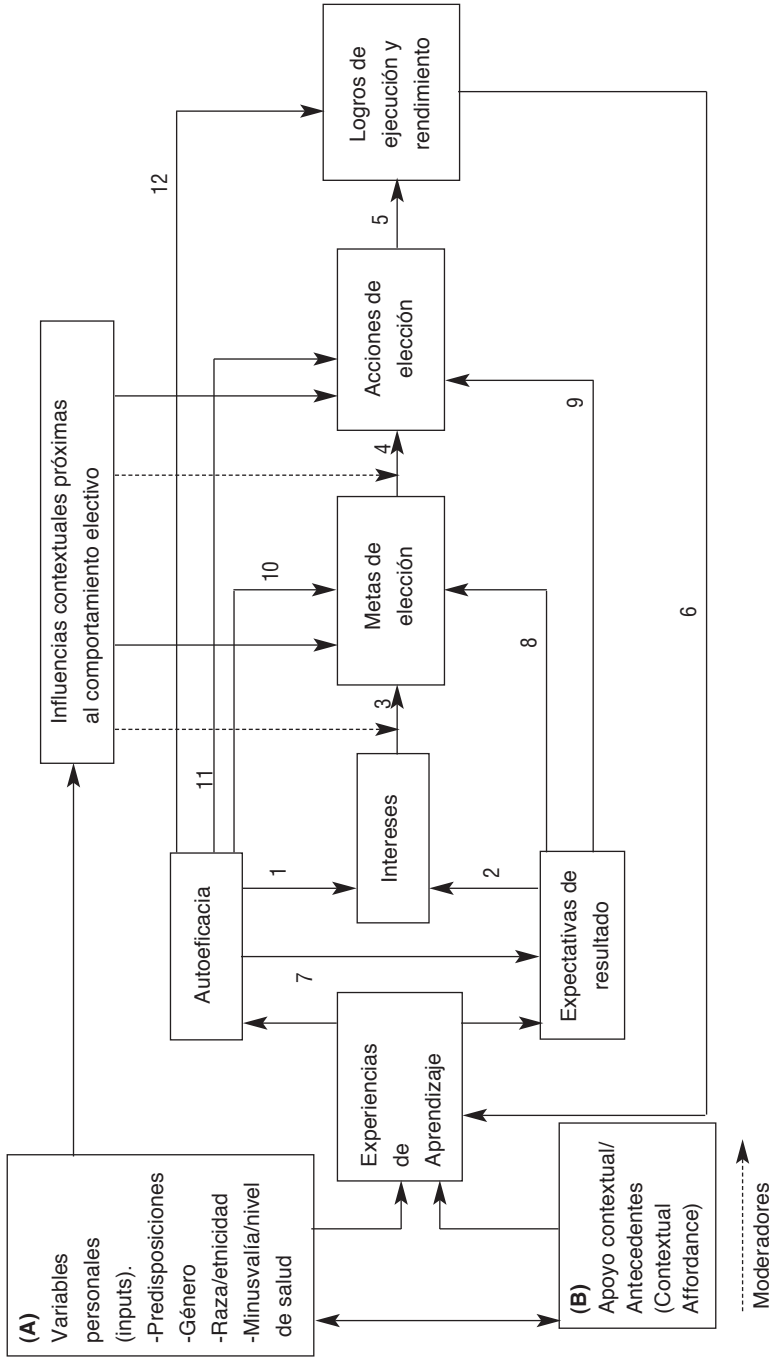


Figura 2.1. Modelo social-cognitivo de desarrollo de carrera

aprendizaje, la matemática, la computación, la escritura, las conductas de prevención de enfermedades de transmisión sexual, el manejo de la tentación de beber y fumar, la enseñanza y el aprendizaje de idiomas, varias de las cuales pueden consultarse en la página mencionada. Bandura (2001) elaboró una monografía para orientar la construcción y análisis psicométrico de este tipo de escalas, la cual es de consulta indispensable para investigadores interesados en la medición de la autoeficacia.

Como ya señaláramos, en nuestro medio Pérez (2001) construyó el Inventario de Autoeficacia para Inteligencias Múltiples (IAMI), con fines de orientación vocacional, que evalúa la seguridad percibida de los adolescentes para realizar exitosamente actividades asociadas con las ocho inteligencias múltiples propuestas por Gardner (1999). El IAMI incluye 8 escalas obtenidas por análisis factorial (Lingüística, por ejemplo) y 64 ítems ("Resolver problemas numéricos", por ejemplo). El usuario de la prueba debe responder utilizando un formato de 10 alternativas, desde (1) "no puedo realizar esa actividad" a (10) "totalmente seguro de poder realizar exitosamente esa actividad". Este inventario está incluido en un Sistema de Orientación Vocacional Informatizado (Fogliatto y Pérez, 2003) y se ha obtenido evidencia favorable de su confiabilidad y validez, esta última respecto de criterios de rendimiento académico y metas de elección de carrera.

Un concepto relacionado con el de autoeficacia es el de autoestima, o autovaloración, que la persona realiza acerca de sí misma. La autoeficacia es una dimensión específica y cognitiva del autoconcepto, así como la autoestima es una dimensión global y valorativa del mismo. En efecto, uno puede valorarse mucho a sí mismo (autoestima elevada) pero no sentirse capaz de realizar una actividad específica (autoeficacia disminuida en algún dominio) y viceversa. En nuestro medio, Grasso (1984) desarrolló una escala para medir la autoestima en ancianos. Este instrumento comprende 15 ítems cuidadosamente elaborados ("Ahora ya no sirvo para nada", por ejemplo) que se responden utilizando una escala *likert* de cuatro posiciones ("Muy de acuerdo", "De acuerdo", "En desacuerdo", "Muy en desacuerdo"). Una validación preliminar de la escala demostró que sus puntuaciones permiten discriminar entre una muestra de an-

cianos internados y otra de individuos más independientes que participan en actividades recreativas en un club de adultos mayores.

Inventarios de intereses vocacionales

Los intereses vocacionales han sido definidos como perfiles de agrados y aversiones respecto a actividades relacionadas con carreras y ocupaciones (Lent, Brown y Hackett, 1994). La problemática de los intereses es de especial utilidad para los investigadores del comportamiento vocacional. Un conocimiento adecuado de esta dimensión de la motivación permite predecir el monto de satisfacción que una persona experimentará en el desempeño de una ocupación. Los intereses se relacionan también significativamente con la estabilidad y el compromiso de los individuos en sus carreras y ocupaciones. Otros investigadores han comparado el peso relativo de los intereses vocacionales en relación con otras variables psicológicas (habilidades, rasgos de personalidad), verificando que los intereses reciben gran consideración por parte del individuo en situaciones de elección de carrera (Holland, 1997).

Los inventarios de intereses son los instrumentos más populares en un contexto de orientación para la elección de carrera, según se desprende de encuestas realizadas en los Estados Unidos, donde instrumentos como el Strong Campbell Interest Inventory (Campbell y Hansen, 1981) son empleados por casi el 90% de los orientadores (Hood y Johnson, 2002). Se los ha definido como una serie de ítems en los que se solicita a los individuos que indiquen sus preferencias vocacionales, a partir de lo cual se pueden obtener puntuaciones finales que representan un perfil de intereses (Cronbach, 1998).

Se coincide en señalar que estos instrumentos deben usarse para seleccionar metas vocacionales, confirmar elecciones previas, descubrir campos de actividad laboral, incrementar el autoconocimiento y encontrar ocupaciones que proporcionen satisfacción (Cronbach, 1998; Hood y Jonhson, 2002). Es claro que los inventarios de intereses poco nos dicen respecto al éxito académico u ocupacional que podrá alcanzar una persona, pero nos

ayudan a identificar carreras u ocupaciones donde puede encontrar satisfacción.

Debe evitarse la práctica profesional de usar los inventarios de intereses para orientar de manera específica a los estudiantes, puesto que éstos necesitan considerar, en el proceso de toma de decisiones de carrera, variables igualmente relevantes y, además, reunir experiencia exploratoria sobre carreras y ocupaciones (Hood y Johnson, 2002). En general, se recomienda confiar en los resultados de estos instrumentos a partir de los 15-17 años, aproximadamente, puesto que se ha verificado que las puntuaciones de los inventarios de intereses son bastante estables a partir de esa edad.

El paradigma más influyente en el dominio de la medición de los intereses vocacionales es el formulado por Holland (1997). La teoría de Holland es un modelo de congruencia entre los intereses y habilidades de una persona, por un lado, y los factores inherentes a su ambiente, por otro. Según este modelo teórico, existen seis tipos de personalidad: Realista, Investigador, Artista, Social, Emprendedor y Convencional (RIASEC), los que a su vez determinan seis patrones análogos de intereses y de habilidades percibidas. El desarrollo de estos tipos depende de una compleja serie de acontecimientos familiares, orientaciones personales iniciales, preferencias ocupacionales e interacciones con contextos ambientales específicos. Los ambientes en los que viven y trabajan las personas pueden también caracterizarse, de acuerdo a su semejanza, con seis modelos que se corresponden con los seis tipos de personalidad anteriormente mencionados.

Los inventarios de intereses vocacionales más populares son el Self-Directed Search (Holland, 1994), el Inventario de Strong-Campbell (Campbell y Hansen, 1981) y el Registro de Preferencias Kuder (Kuder y Zitowsky (1991). Más allá de sus diferencias (Kuder obtuvo sus escalas por análisis factorial y emplea ítems de elección forzosa, el inventario Strong posee claves ocupacionales formadas por la comparación de personas satisfechas en una ocupación con respuestas de la muestra de estandarización), todos utilizan el modelo teórico RIASEC para interpretar sus resultados, lo cual permite una convergencia conceptual impensable en otros dominios de la psicología. Una iniciativa interesante es el Test Visual de Intereses Profesionales (Tetreau y

Trahan, 1986), desarrollado por investigadores canadienses y basado también en el modelo de Holland, pero que utiliza 80 fotografías en color ilustrando actividades laborales en lugar de ítems verbales, con el fin de atenuar los problemas de sesgo cultural que generan estos últimos reactivos.

Recientemente se construyó un nuevo inventario de intereses, con promisorias perspectivas. En efecto, el Personal Globe Inventory (Tracey, 2002) incluye ocho escalas básicas de intereses (Servicio, Relaciones Públicas, Asistencia, Arte, Ciencias de la Vida, Mecánica, Tecnología y Negocios) semejantes al modelo RIASEC, aunque con mayor especificidad. La innovación quizá más importante que introduce es su discriminación entre profesiones de alto y bajo prestigio social, asociadas a sus ocho escalas. Esto permite que el inventario pueda ser empleado para brindar orientación a trabajadores poco calificados, y no solamente a estudiantes que aspiran a continuar una carrera superior. Los datos preliminares demuestran fuertes propiedades psicométricas de las escalas de este test.

Si bien existe evidencia preliminar de la influencia genética sobre los intereses vocacionales, existen interrogantes básicos que deberán ser esclarecidos en el futuro, tales como ¿cuáles son las bases neurobiológicas de los intereses vocacionales? o ¿en qué medida pueden diferenciarse de otros constructos relacionados (rasgos de personalidad o actitudes, por ejemplo)? Para poseer una teoría científica de los intereses vocacionales debe contarse con teorías explicativas y universales. En efecto, el modelo RIASEC es preponderantemente descriptivo y no ha logrado replicarse bien en algunos contextos culturales diferentes del occidental.

Hay una gran variedad de tests de intereses vocacionales pero se presentan dificultades considerables cuando se emplean de modo transcultural. Uno de los obstáculos más significativos en la traducción y adaptación de tests verbales son los problemas de lenguaje. En este sentido, las traducciones libres pueden traicionar las intenciones originales del autor, y las literales, por los problemas de equivalencia semántica y la diferente frecuencia de uso de las palabras en lenguas diversas, no alcanzar a expresar con precisión los significados de los ítems en sus versiones originales.

También debe considerarse el papel de los factores culturales que pueden falsear de algún modo los resultados de tests cuando se emplean en otras culturas (véase capítulo 7, “Adaptación de tests a otras culturas”). Es frecuente encontrar, en los inventarios de intereses, ítems que mencionan actividades que en las culturas de origen tienen una popularidad que no poseen en otros contextos; jugar béisbol, por ejemplo, tiene un significado diferente en aquellos países donde es un deporte poco practicado. Algunos ítems mencionan títulos u ocupaciones que son familiares en el país de origen del inventario y, en cambio, resultan extraños para los ciudadanos de otras naciones (Fogliatto, 1991).

Estos problemas indican con claridad los riesgos de emplear de un modo acrítico los tests construidos en otras culturas. Fogliatto planteó la necesidad de construir un cuestionario de intereses de características locales y más adecuadas a las preferencias, actividades educacionales y laborales, así como al lenguaje habitual de los jóvenes de nuestro medio. Este instrumento es su Cuestionario de Intereses Profesionales (CIP) (Fogliatto, 1991).

Tabla 2.2. Muestra de ítems del Cuestionario de Intereses Profesionales Revisado (CIP-R)

	D	I	A
1. Aprender estilos de pintura artística.			
2. Cantar en coros.			
3. Trabajar en estudios jurídicos.			
4. Trabajar con calculadoras.			
5. Aprender a interpretar radiografías.			
6. Enseñar a niños.			
7. Asesorar sobre el cuidado de plantas.			

La última versión del Cuestionario de Intereses Profesionales (CIP-R) es asistida por computadora y se integra al Sistema de Orientación Vocacional Informatizado (Fogliatto y Pérez,

2003) que incluye también el IAMI (Pérez, 2001), así como un banco de información académica sobre carreras y especialidades educativas del secundario. El CIP-R comprende 15 escalas (Cálculo, Asistencial, Musical, Artística, entre otras) y 114 ítems que describen actividades académicas o laborales. La persona debe responder utilizando tres alternativas de respuesta: Agrado, Indiferencia o Desagrado a cada uno de los ítems, por ejemplo "Construir puentes". El CIP-R posee buenas propiedades de confiabilidad y validez de criterio con respecto a metas de elección de carrera. Un aspecto criticable de este inventario es que su construcción ha sido empírica, basándose exclusivamente en el análisis factorial exploratorio de ítems relacionados con carreras y no en una teoría explícita de los intereses vocacionales. Por consiguiente, es dificultoso relacionar sus escalas con los constructos de teorías reconocidas, tales como el modelo RIASEC (Holland, 1997).

Escalas de actitudes

Las actitudes se refieren a predisposiciones aprendidas para responder positiva o negativamente ante objetos sociales particulares, es decir, tipos de personas, instituciones sociales o situaciones (Aiken, 2003). Para Padua (1979), las actitudes son tendencias individuales a reaccionar, positiva o negativamente, frente a un valor social.

Desde el punto de vista conceptual es díficil diferenciar las actitudes de los intereses (Anastasi y Urbina, 1998). Al respecto, piénsese cómo podría distinguirse una escala de intereses por el cálculo y otra de actitudes ante la matemática, por ejemplo. Probablemente la diferencia esencial entre estos dos constructos radique en el área de la psicología donde se apliquen. En efecto, los inventarios de intereses miden patrones de preferencias (y rechazos) por áreas de conocimiento o trabajo y son utilizados casi exclusivamente por los orientadores vocacionales; las escalas de actitudes, en cambio, generalmente miden patrones de preferencias (y rechazos) por creencias e ideologías y por consiguiente son más empleadas en la psicología social o política.

Las escalas de actitudes surgieron como una preocupación de los investigadores frente a la problemática de la aceptación social de grupos (y creencias) diferentes (Anastasi y Urbina, 1998). La primera escala de actitud fue la de distancia social (Bogardus, 1925) donde los examinados clasificaban varios grupos raciales y religiosos en función de su aceptación. Fue notable la producción de escalas de actitudes porteriormente a la Segunda Guerra Mundial, tales como la famosa escala para medir el autoritarismo (Adorno, Frenkel-Brunswik, Levinson y Sanford, 1950).

La medición de actitudes, además de tener múltiples aplicaciones, también posee distintas variantes, aunque en la actualidad predominan las escalas tipo *likert* (Likert, 1932). Estas escalas se caracterizan por presentar afirmaciones (ítems) que deben responderse empleando una escala de 5 a 7 alternativas que indican el acuerdo del examinado con el contenido enunciado en cada ítem. Si bien, tradicionalmente, la construcción de escalas de actitud se caracterizó por el uso de procedimientos específicos, en la literatura actual son indicados los lineamientos generales de construcción de tests, que expondremos en el capítulo 6.

El desarrollo de escalas de actitudes consiste, inicialmente, en elaborar un conjunto de ítems relativos a la dimensión que se pretende medir y asignar números a las diversas alternativas de respuesta a esos ítems. Esos valores numéricos reflejan la intensidad de la actitud, positiva o negativa, que posee un sujeto frente a un objeto determinado. Los procedimientos de determinación de la confiabilidad y validez de las escalas de actitudes tampoco difieren de los utilizados en los otros tipos de tests.

Aiken (2003) construyó una escala *likert* de actitud ante la matemática, luego adaptada por Murat (1984) para nuestro medio. Una escala frecuentemente citada en la investigación contemporánea es la de roles sexuales de Bem (1974) que mide actitudes hacia la masculinidad y la feminidad (como estereotipos sociales). Cada ítem describe algunas características personales típicas de los géneros, y la persona que responde debe indicar su grado de acuerdo empleando una escala *likert* de siete puntos (muy de acuerdo, bastante de acuerdo, algo de acuerdo, ni acuerdo ni desacuerdo, algo en desacuerdo, bastante en desacuerdo y muy en desacuerdo).

Tornimbeni y González (1997) elaboraron para nuestro país una escala de actitud hacia la investigación, cuyos ítems son del siguiente tipo.

- Recién al finalizar mi carrera voy a pensar en la posibilidad de convertirme en investigador.

Esta escala posee 5 alternativas de respuesta: muy de acuerdo, acuerdo, ni acuerdo-ni desacuerdo, desacuerdo y muy en desacuerdo.

Inventarios de rasgos de personalidad

Otros tests de uso frecuente en la psicología contemporánea son aquéllos construidos para medir rasgos de personalidad. Aun cuando el término "personalidad" sea empleado en diferentes acepciones y carezca de un sentido unívoco, la mayor parte de las definiciones coinciden en que hace referencia a las tendencias afectivas básicas de una persona. Estas disposiciones le confieren relativa estabilidad al comportamiento individual, más allá de las lógicas variaciones que resultan de la adaptación a diferentes contextos y situaciones.

Según Nunnally (1991) el estudio total de la personalidad se centra en dos grandes problemas:

1) Cuáles son los rasgos dominantes de una persona en un momento determinado de su historia personal.
2) Qué factores determinan ese perfil de personalidad.

La medición de la personalidad atañe principalmente al primer punto, y su propósito principal es describir a los individuos sobre la base de sus rasgos de personalidad predominantes. El segundo punto se relaciona con la herencia y la experiencia, ya que para explicar el desarrollo de la personalidad de un individuo se debe recurrir a la genética del comportamiento y a las teorías del aprendizaje.

En algunas de las teorías contemporáneas, tales como la de los cinco grandes factores (Costa y Mc Crae, 1999), los rasgos de

la personalidad se entienden como hereditarios en gran proporción y, por consiguiente, bastante asimilables al concepto de "temperamento" o "naturaleza emocional" de las personas (Carver y Scheier, 1996). Algunas de las orientaciones temperamentales básicas, tales como emocionalidad positiva (asimilable a Extraversión y Amabilidad) y negativa (asimilable a Neuroticismo), pueden distinguirse ya en la primera infancia (Tellegen, 1988). La investigación actual en genética del comportamiento (Plomin y colaboradores, 2002) apoya este condicionamiento hereditario de la reactividad emocional de las personas, aunque admite que el entorno familiar también explica parte de la variabilidad de esa variable. Recientemente se ha sugerido que el incremento en la actividad social, el ejercicio físico y las técnicas de relajación pueden modificar algunas tendencias emocionales básicas de las personas (Lent, 2004).

Los rasgos de personalidad se relacionan con la conducta típica de las personas en su vida cotidiana, tales como el nivel de ansiedad o de amabilidad. Existe un buen número de estrategias diferentes para medirlos, aunque en los últimos años se utilizan preferentemente los inventarios autodescriptivos o de autoinforme (Casullo y cols., 1994). Un ítem típico de este tipo de inventarios puede ser como el siguiente: "Me agradan las reuniones sociales."

Las opciones de respuesta a ítems como el anterior pueden ir desde un formato dicotómico ("Sí-No" o "Verdadero-Falso") a uno de tipo *likert*. Actualmente se recomienda incluir varias alternativas de respuestas con la finalidad de mejorar la variabilidad de las respuestas y, por consiguiente, la confiabilidad y validez de los tests (Pajares, Hartley y Valiante, 2001).

Los inventarios de rasgos de personalidad se utilizan en ámbitos tan diversos como la clínica psicológica, la psicología ocupacional y la investigación. Como afirmamos antes, uno de los principales inconvenientes de los autoinformes es la posibilidad de que los sujetos falseen sus respuestas para dar una impresión socialmente aceptable (Anastasi y Urbina, 1998).

Pueden distinguirse dos tipos de inventarios de personalidad: los que evalúan rasgos psicopatológicos y los que miden rasgos de la personalidad "normal". Entre los primeros, de uso preferentemente clínico, uno de los más utilizados es el Inven-

tario Multifásico de Personalidad de Minnesota (MMPI), elaborado en la década de 1940 para diagnosticar trastornos psicológicos. Los 500 ítems del MMPI incluyen una amplia variedad de contenidos y comprenden áreas como actitudes sexuales, educación, ocupación, familia, salud, síntomas psicosomáticos, manifestaciones neuróticas y psicóticas de la conducta, etc. En su versión original permite obtener puntuaciones en diferentes escalas clínicas relacionadas con distintas categorías de la psicopatología clásica (histeria, hipocondría, por ejemplo).

El MMPI-2 (Butcher, Dahlstrom, Graham, Telegen y Kaemmer, 1989) es una versión revisada y actualizada que incluye nuevos ítems, escalas adicionales y baremos actualizados. La estructura interna del MMPI (constructos medidos por las diferentes escalas del test) ha sido cuestionada por los análisis factoriales realizados, que tienden a identificar dos factores (afectividad positiva y negativa) consistentes (Kaplan y Saccuzzo, 2006). No obstante, el MMPI es uno de los tests más populares e investigados del mundo y, en los Estados Unidos, es aceptado como evidencia adicional en un proceso judicial.

Otros inventarios de personalidad de uso clínico miden un trastorno psicológico específico, como el Inventario de Depresión BDI-II (Beck, Steer y Brown, 1996) o el Test de Ansiedad Rasgo-Estado (Spielberger, 1983), entre otros numerosos instrumentos de este tipo.

Entre los inventarios usados para evaluar rasgos de personalidad en personas sin trastornos psicológicos severos, los más populares son el 16PF-5 (Russell y Karol, 2000), el EPQ-R (Eysenck y Eysenck, 1997), y el NEO-PI-R (Costa y Mc Crae, 1999).

La teoría de los cinco grandes factores (Norman, 1963; Costa y Mc Crae, 1999) es predominante en la construcción de los inventarios de personalidad elaborados para medir predisposiciones no patológicas. Esta teoría postula cinco dimensiones afectivas básicas en las cuales diferimos los seres humanos: Estabilidad Emocional, Extroversión, Apertura, Responsabilidad y Amabilidad. El volumen de investigación acerca de este modelo es abrumador, aunque como en el caso de la inteligencia, existen varias teorías alternativas y competidoras, como la teoría PEN de Eysenck (1981), que propone tres factores (Neuroticismo, Psicoticismo y Extroversión) en lugar de cinco.

Pueden establecerse relaciones entre ambas teorías, puesto que dos constructos son perfectamente asimilables: Extroversión y Neuroticismo (el polo negativo de Estabilidad Emocional) y el tercer factor de la teoría PEN, Psicoticismo (también denominado Impulsividad), se relaciona con Responsabilidad y Amabilidad de manera inversa. Por otra parte, el factor de Apertura (o Intelecto) de la teoría de los cinco grandes factores (Costa y Mc Crae, 1999) no es reconocido por Eynseck (1981) como un factor de personalidad. En síntesis, y tal como acontece con la inteligencia, el dominio de la personalidad es altamente controversial. La revista *Personality and Individual Differences* es una de las mejores fuentes de consulta sobre la medición y teoría de la personalidad.

El NEO-PI-R (Costa y Mc Crae, 1999) mide los cinco grandes factores y 30 facetas específicas que permiten una mayor discriminación en la medición de la personalidad. El NEO-PI-R se emplea en diferentes áreas de la psicología aplicada (en especial en el ámbito laboral) y en la investigación. También existe una versión abreviada de este inventario, el NEO-FFI, que mide solamente los cinco factores principales sin las respectivas facetas.

Uno de los principales investigadores del modelo de los cinco factores, Goldberg (1999), diseñó un banco internacional de ítems (*international pool items personality*, IPIP), a disposición en la Web para los usuarios interesados en utilizar, investigar o construir inventarios de medición de la personalidad (www.ipip.org).

Los inventarios que miden rasgos de personalidad "normales", tales como el NEO en sus diferentes versiones (Costa y Mc Crae, 1999) y el inventario 16PF-5 (Russell y Karol, 2000), se emplean crecientemente en psicología ocupacional y educacional, aunque también en contextos clínicos, en especial para diseñar programas de intervención preventivos, relacionados con el manejo de la afectividad y los vínculos interpersonales. Varias investigaciones han demostrado que los factores Responsabilidad y Apertura, en particular, son predictivos del rendimiento académico y ocupacional (Tokar, Fisher y Subich, 1998). Por su parte, Extraversión y Neuroticismo son factores asociados con la satisfacción en el empleo y el bienestar psicológico general (Lent, 2004).

Otra estrategia de medición de la personalidad son las denominadas técnicas proyectivas, que emplean estímulos (ítems) ambiguos ante los cuales se espera que los sujetos "proyecten" sus sentimientos, deseos y emociones. Las técnicas proyectivas poseen varias limitaciones que aconsejan su empleo como método de investigación más que de diagnóstico. Éstas comprenden: pobre confiabilidad, baja validez, carencia de un método objetivo para puntuar e influencias contextuales sobre los puntajes (Kline, 2000).

El test proyectivo más conocido es el Psicodiagnóstico de Rorschach (1921) ya mencionado en el apartado histórico de la primera parte de este texto. Incluye diez láminas (manchas de tinta simétricas) y las características de las respuestas son interpretadas por medio de parámetros preestablecidos, tales como atender a los detalles o a la figura global; o responder preferentemente al color o la forma. En los últimos años se han realizado intentos por dotar de mayor estandarización a las condiciones de administración, puntuación e interpretación de sus resultados. Exner (1993) elaboró un sistema muy aceptado que ha mejorado la confiabilidad de las puntuaciones del Rorschach, aunque la evidencia es mixta respecto a su validez (Hogan, 2004).

Inventarios de habilidades sociales

Finalmente, otro desarrollo psicométrico contemporáneo es el de la medición de las Habilidades Sociales (HHSS), constructo proveniente de la psicología cognitivo-comportamental, y de gran relevancia en la evaluación clínica, educativa y ocupacional. El término "habilidades sociales" se introduce en la literatura en la segunda mitad de los años setenta, y a partir de la década siguiente se observa un incremento de la evaluación de habilidades sociales en diferentes ámbitos, tales como la psicología educativa, clínica y ocupacional (Mac Combs y Branan, 1990).

Las HHSS han sido definidas como el conjunto de conductas que favorecen el desarrollo social de la persona y por medio de las cuales ésta expresa sus sentimientos, actitudes, deseos, opiniones o derechos de un modo adecuado a la situación, respetando la expresión de esas conductas en los demás. Para Kelly (1987), el concepto de HHSS incluye diferentes subcompetencias tales como habilidades conversacionales, habilidades heterosociales de concertación de citas, habilidades para entrevistas de trabajo, oposición asertiva y aceptación asertiva.

Se han construido varias medidas de autoinforme de las HHSS, tales como el Inventario de Asertividad de Rathus (en Kelly, 1987) y la Escala Multidimensional de Expresión Social (Caballo, 1987). Si bien algunos instrumentos han sido adaptados a nuestro medio, carecemos de inventarios locales de evaluación del constructo, por lo cual la elaboración de este tipo de tests constituye un área de interés científico y aplicado en la región.

Finalmente, cabe señalar que, en estos últimos años, se han construido tests que no sólo contemplan características intrapsicológicas (cognitivas o afectivas) sino que también miden aspectos relacionados con los diferentes ambientes en los cuales se desenvuelve el individuo. De este modo, existen tests para medir dimensiones del ambiente social, escolar u ocupacional (Kaplan y Saccuzzo, 2006). Estos desarrollos son muy interesantes puesto que reconocen al comportamiento del ser humano como una función de su sistema nervioso (incluidos los componentes psicológicos), la sociedad y la interacción entre ambos factores, tal como ha sido remarcado por varios autores (Bandura, 1997; Bunge y Ardila, 2002).

Para finalizar, una sucinta referencia a una destacada investigadora argentina (al igual que la Dra. Cortada de Kohan y la Dra. Casullo, mencionadas anteriormente) que trabaja asiduamente en la construcción y adaptación de tests de respuesta típica: la Dra. Richaud de Minzi. Entre sus contribuciones en este ámbito pueden citarse la construcción de escalas para medir estilos de afrontamiento en niños y estilos parentales (Richaud de Minzi, 2005), así como diversas adaptaciones de tests de personalidad, como el Inventario Beck y el NEO, ya mencionados anteriormente.

Segunda Parte

NORMAS TÉCNICAS

INTRODUCCIÓN

Para que un test pueda ser utilizado responsablemente es necesario que cumpla con determinados estándares técnicos. Hasta mediados del siglo pasado los instrumentos de medición en psicología fueron aplicados con escaso control de su calidad; es más, podría decirse que se administraron e interpretaron sin una clara demostración de su utilidad para los fines propuestos, ni de sus límites o alcances. Esto ocasionó innumerables críticas y, en muchos casos, un abierto rechazo social a la utilización de tests. Como consecuencia de estos cuestionamientos surgió una corriente de revisión y análisis de la fundamentación científica de las pruebas.

En los Estados Unidos se publicaron documentos generados por organizaciones especializadas (American Psychological Association, American Educational Research Association), cuya meta esencial fue establecer los requisitos técnicos mínimos que debían reunir los tests utilizados en el ámbito de la psicología y la educación. En 1966 se publicaron las "Normas técnicas para tests psicológicos y educativos" y en 1999 apareció la última revisión de las mismas, que incluye modificaciones importantes, particularmente en la concepción de validez de los tests. Los profesionales usuarios de tests deben conocer estas normas y ajustarse estrictamente a ellas en lo concerniente a la administración, validación e interpretación de los resultados de estos instrumentos. Debido a su importancia, las normas técnicas de los tests serán analizadas con detenimiento en los capítulos siguientes.

3
CONFIABILIDAD

Fabián O. Olaz

3.1. Introducción

Como se afirmó en el capítulo inicial, la medición psicológica parte de ciertos supuestos fundamentales. Uno de éstos expresa que el resultado de la medición es un valor observado que no coincide con el valor verdadero y, en consecuencia, siempre se mide con un margen de error. Este valor verdadero es un valor teórico, un concepto matemático. En términos matemáticos, este valor es la esperanza matemática de la puntuación observada y podría pensarse como la media de las puntuaciones observadas obtenida de infinitas administraciones de un instrumento dado a una persona (Muñiz, 2001). Tomando en consideración este supuesto, se puede inferir que cuanto mayor sea el error, menos confiables serán los resultados obtenidos en el proceso de medición. Es importante considerar que con el término "error" nos referimos a cualquier variación de las puntuaciones de un test que no sea asimilable a las variaciones en la magnitud del rasgo que está siendo evaluado (por ejemplo, los cambios en la autoeficacia de una persona entre una medición y otra). Siempre que medimos repetidamente un fenómeno, sea éste de naturaleza física o social, es inevitable una cierta dosis de error, debido a imprecisiones del instrumento o a la influencia de las posibles fuentes de variación de las puntuaciones de un test, que analizaremos más abajo.

De este modo, si medimos repetidamente la longitud de un objeto determinado utilizando una regla metálica, probablemente obtengamos resultados casi idénticos en todas las ocasio-

nes. En este caso, los datos obtenidos tienen un nivel elevado de consistencia o replicabilidad, y si encontramos variaciones entre una medición y otra, podemos inferir que se deben a cambios en el objeto medido. Si las mismas mediciones se efectúan con una cinta elástica, al medir repetidamente el mismo objeto puede obtenerse una distribución de valores numéricos con una cierta dispersión y, por consiguiente, los datos tendrán un nivel de consistencia más bajo que en el caso anterior. Esta dispersión de los valores obtenidos durante mediciones repetidas, bajo condiciones similares, se relaciona con el concepto de confiabilidad. En síntesis, cuanto mayor es la variabilidad entre las medidas del mismo fenómeno en repetidas ocasiones tanto menor es la confiabilidad, y cuanto menor es la variabilidad mayor la confiabilidad.

En el dominio de los tests psicológicos esta variabilidad es mayor que en la medición de los fenómenos físicos, debido a las características muy dinámicas del objeto de medición (el comportamiento humano) y la mayor cantidad de fuentes de error que pueden afectar las puntuaciones, en comparación con otros dominios del conocimiento. En una persona las diferencias en el desempeño en un test en diversas ocasiones pueden originarse en una motivación diferente en una y otra situación de prueba, distintos niveles de fatiga, una mayor familiaridad con el contenido del test y una variedad de factores similares. Por estas razones, el puntaje de una persona en un test psicológico nunca será perfectamente consistente de una ocasión a la próxima aun en el caso de que se la evalúe con una misma prueba.

3.2. El concepto de confiabilidad en la teoría clásica de los tests

En la actualidad coexisten dos teorías generales de los tests, la teoría clásica de los tests y la de respuesta al ítem. Los fundamentos de ambos modelos teóricos se desarrollarán en el último capítulo.

La hipótesis fundamental de la teoría clásica de los tests (TCT) es que la puntuación observada de una persona en un test es una función de dos componentes: su puntaje verdadero

(que es inobservable) y el error de medición implícito en toda medición.

El postulado esencial de la TCT se expresa como:

$$Oi = Vi + Ei,$$

Esto es, la puntuación observada de un individuo es igual a la puntuación verdadera más el error. En el plano teórico, la puntuación verdadera puede entenderse como la media de las puntuaciones obtenidas por una persona en infinitas aplicaciones de un test (en diferentes momentos y condiciones), asumiendo que la forma de distribución de esas infinitas puntuaciones se aproxima a la normal.

La puntuación de error, por otra parte, es la suma de todos aquellos factores aleatorios que influyen y afectan el registro de los datos, introduciendo inconsistencia en el proceso y alejando la puntuación observada de la puntuación verdadera. Aun en ciencias naturales, donde se cuenta con instrumentos más precisos, existe esa posibilidad de error. En la medida en que controlemos las fuentes de error de una medición, más se acercará la puntuación observada (empírica) a la puntuación verdadera (teórica).

La confiabilidad puede entenderse como la exactitud o precisión de una medición, o el grado en el cual las puntuaciones de un test están libres de esos errores de medición. Esta exactitud o precisión de las puntuaciones permite que éstas se mantengan constantes en diferentes circunstancias.

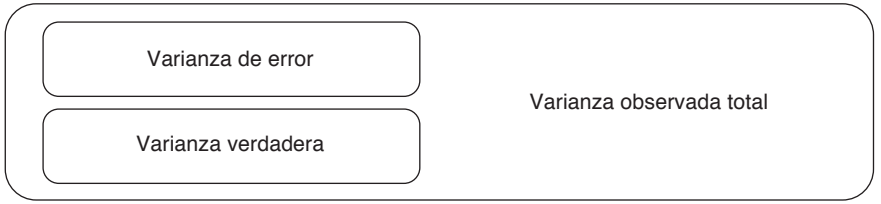
> La confiabilidad significa la consistencia entre los puntajes de un test obtenidos por los mismos individuos en distintas ocasiones o entre diferentes conjuntos de ítems equivalentes (APA, 1999).

Obviamente, debemos distinguir entre la falta de consistencia debido a un cambio real en el rasgo medido y las fluctuaciones azarosas de las puntuaciones ocasionadas por cambios en circunstancias personales u otras que afecten la situación de evaluación. Los cambios reales en el atributo medido no son

fuente de falta de confiabilidad, mientras que todo cambio debido a factores externos que se presenta en forma no sistemática afecta la confiabilidad de las puntuaciones.

Es importante diferenciar entre errores sistemáticos y aleatorios. Un error sistemático es aquel que genera constantemente una puntuación elevada o baja en una persona al margen de los cambios que puedan darse en el rasgo medido por el test. Este tipo de error se denomina sesgo de medición, que se define como toda aquella fuente de variación que sistemáticamente afecta a las puntuaciones de un determinado grupo que está siendo evaluado por un test, ya sea elevando o disminuyendo las mismas. De esta manera, estos sesgos afectan directamente las inferencias o interpretaciones que podemos realizar a partir de esos puntajes. Un instrumento que mide aptitudes para la matemática puede estar construido de tal manera que sus ítems sean poco comprensibles para un grupo social o etario determinado, y esto podría ocasionar que ese grupo obtenga puntuaciones más bajas en el test independientemente de sus aptitudes matemáticas. Si bien los sesgos están habitualmente presentes en todo proceso de medición, este tipo de error es predecible y los diferentes procedimientos que se exponen en el capítulo 4 sobre Validez apuntan a controlarlos. Por otra parte, los errores aleatorios que afectan la precisión o consistencia de las medidas realizadas son impredecibles y forman parte de todo proceso de medición. El control de estas fuentes de error se relaciona con los métodos utilizados para verificar la confiabilidad.

En general, la confiabilidad se verifica mediante un coeficiente de correlación entre las medidas repetidas de un fenómeno. Para entender la lógica de la utilización de este coeficiente cabe realizar algunas observaciones previas. La puntuación verdadera también puede expresarse en términos de varianza de las puntuaciones de los tests. Recuérdese que la varianza indica la dispersión promedio de los valores (en este caso los puntajes de tests) alrededor de la media de un grupo de observaciones (más precisamente, el promedio de los cuadrados de la diferencia entre cada valor y la media). En la teoría clásica de los tests, la confiabilidad puede expresarse como la proporción de varianza observada de las puntuaciones de tests que se debe a la varianza verdadera (la variabilidad de la puntuación verdadera), es decir, a la varianza del rasgo que se está evaluando y no a otros factores. Puede representarse lo afirmado anteriormente de la siguiente manera:



Así considerada, la confiabilidad se suele simbolizar como la razón de la varianza verdadera y la varianza observada:

$$r_{xx©} = \frac{\sigma_v^2}{\sigma_o^2}$$

Donde:

$r_{xx©}$ = coeficiente de correlación

σ_v^2 = varianza verdadera

σ_o^2 = varianza observada

Los diferentes métodos utilizados para evaluar la confiabilidad utilizan el coeficiente de correlación como estadístico fundamental. De este modo, un coeficiente de 0,80 sugiere que un 64% de la varianza observada es asimilable a la variabilidad de la puntuación verdadera, con un margen de error del 36%. Los diferentes métodos que se expondrán más adelante están diseñados para estimar la cuantía de error presente en las puntuaciones de un test determinado.

Como puede inferirse, la confiabilidad (como la validez) no es una característica del test en sí mismo, según la teoría clásica de los tests, sino una propiedad de las puntuaciones del test cuando éste se administra a una muestra específica y bajo condiciones particulares (APA, 1999). Ésta es una de las principales limitaciones de la TCT. Como veremos en el capítulo 8, esta

limitación es atenuada en los tests que se construyen utilizando las herramientas metodológicas y los supuestos de la teoría de respuesta al ítem.

3.3. Principales factores que afectan la confiabilidad

Existen múltiples factores que afectan la confiabilidad de las puntuaciones obtenidas mediante la aplicación de un test. La APA (1999) los clasifica en factores internos (fluctuaciones en el nivel de atención del examinado, por ejemplo) o externos (diferentes criterios de corrección de un test, según evaluadores distintos, por ejemplo). A continuación revisamos estos factores de error, agrupándolos en cuatro categorías de acuerdo al criterio propuesto por Hogan (2004). Esta clasificación no pretende ser exhaustiva, sino solamente destacar aquellos errores fundamentales que afectan la confiabilidad de las puntuaciones.

Contenido del test

Al construir un test debe tenerse en cuenta que la elección de los ítems, los materiales a través de los cuales estimulamos la respuesta del individuo, y la formulación de las consignas pueden ser una fuente de error aleatorio en la medición. Si se desea estimar el puntaje de un individuo en un cierto dominio, los errores en el muestreo de contenido pueden afectar la confiabilidad.

Si consideramos a un test como un conjunto de indicadores operacionales de un rasgo o dominio de comportamiento, es comprensible que la elección de los mismos (muestra de ítems) pueda constituir una fuente de error. Tanto si el muestreo de contenido es sesgado, como si no es suficientemente extenso, los puntajes resultantes serán poco confiables. Un test de inteligencia con solo 10 ítems representa un ejemplo de muestreo de contenido muy limitado, pues en este caso los puntajes dependerían de la capacidad del individuo con respecto al pequeño conjunto de reactivos utilizados. Es esencial que los ítems de un test estén fuertemente interrelacionados y midan un mismo

constructo (unidimensionalidad), de otro modo se introducirían inconsistencias entre las diferentes partes del test y, por consiguiente, habría fuentes de error en la medición.

En general un número mayor de ítems del mismo tipo dará como resultado puntajes más confiables. Como se desprende de los supuestos del modelo clásico de los tests, si se tiene un test determinado y se aumenta su longitud n veces, con ítems paralelos a los originales, la fiabilidad del nuevo test aumenta. Esto se conoce como profecía de Spearman-Brown. No obstante, este postulado debe tomarse con precaución por dos motivos. En primer lugar, en la literatura psicométrica es habitual que se informen índices de confiabilidad óptimos en tests breves y sólo moderados en tests extensos. En segundo lugar, la confiabilidad de una prueba se relaciona también con la intercorrelación de los ítems, observándose (aun en el caso de que haya pocos ítems) que la confiabilidad será mayor cuanto mayor sea la intercorrelación de los mismos. En este sentido, la calidad de los ítems es un factor más importante que la mera cantidad, y deben evitarse los ítems redundantes.

Algunos ítems no requieren que la persona reconozca la respuesta correcta. Éstos son lo denominados “ítems de reconocimiento” que se emplean en las pruebas de opción múltiple (*multiple choice*). Siempre que un ítem suponga reconocimiento de una o más respuestas correctas, la posibilidad de adivinación desempeña un papel en los puntajes obtenidos en el test. Esta posibilidad está inversamente relacionada con el número de respuestas alternativas para cada ítem; así por ejemplo, en un ítem dicotómico (verdadero - falso; correcto - incorrecto) hay una posibilidad sobre dos de obtener una respuesta correcta por mera adivinación. Por el contrario, en un test de opción múltiple, con varias respuestas posibles para cada ítem, se podrían identificar muy pocas respuestas correctas obtenidas exclusivamente por azar.

En términos generales, los tests son más confiables a medida que aumenta el número de respuestas alternativas dentro de un rango limitado (Cortada de Kohan, 1999). Pajares, Hartley y Valiante (2001) compararon los índices de confiabilidad y validez de dos escalas de autoeficacia para la escritura, una con un formato *likert* clásico (de siete opciones de respuesta) y otra

con diez alternativas de respuesta, encontrándose mejores indicadores psicométricos en esta última forma.

En conclusión, tanto si las puntuaciones de un test son afectadas por el muestreo de contenido o por el uso de una determinada escala de respuesta, las diferencias que se observen en las puntuaciones de las personas evaluadas se deberán no a diferencias reales en el rasgo medido sino a variaciones relacionadas con errores de medición.

Administración

En el momento de administrar un test también pueden introducirse errores que afecten la confiabilidad de los resultados. Por ese motivo es esencial examinar a todos los participantes en condiciones uniformes, estandarizadas. Así, por ejemplo, las condiciones generales del ambiente en que se administran los tests deben ser lo más semejantes que sea posible (iluminación, nivel de ruido o confort del lugar). Si bien la estandarización debe estar prevista claramente en los materiales del test, pueden producirse variaciones en el momento de administración. En los tests de velocidad, por ejemplo, si un administrador es demasiado flexible en la asignación de los límites del tiempo de respuesta, un grupo podría obtener puntuaciones más altas que los que hubiera obtenido con un administrador más respetuoso del tiempo pautado en el manual del test (Nunnally, 1991).

Por consiguiente, siempre es deseable que las instrucciones del test sean lo suficientemente claras y unívocas, para que todos los evaluadores las impartan de la misma manera y presenten los materiales en idéntico orden y forma a todos los examinados. La falta de consistencia en la administración de un test influirá en la estabilidad de las puntuaciones obtenidas por las personas medidas por ese test.

Calificación

En el momento de calificar (puntuar) un test existen otros factores que pueden influir negativamente en la confiabilidad

de las puntuaciones obtenidas. En los tests de opción múltiple pueden cometerse errores cuando la corrección es manual, tales como calificar accidentalmente algunas respuestas correctas como erróneas y viceversa, o realizar mal la sumatoria de los respuestas clave o correctas. La posibilidad de cometer este tipo de errores prácticamente se elimina en los tests computarizados.

No obstante, una de las fuentes principales de no-confiabilidad es que los diferentes evaluadores utilicen criterios distintos de calificación. Este aspecto es crítico en los tests que permiten un mayor margen de subjetividad en la interpretación de los resultados, tal como acontece en algunos subtests de las escalas Wechsler de Inteligencia. Así, por ejemplo, un test verbal puede solicitar a los examinados que den una definición de una palabra (escuela, por ejemplo). La respuesta proporcionada ("lugar al que los estudiantes van para aprender" o "edificio donde vive el maestro", por ejemplo) debe ser calificada con 0, 1 ó 2, según su adecuación. Muchas veces los criterios para considerar "totalmente adecuada" una respuesta no son lo suficientemente claros y esto genera falta de acuerdo entre los calificadores del test, lo que a su vez genera variaciones no sistemáticas en las puntuaciones de las personas en ese test.

De allí que, cuando en el proceso de puntuación de un test interviene de manera importante el criterio del evaluador, puedan presentarse variaciones que disminuyan la confiabilidad de los puntajes.

Factores internos del examinado

Parte de la varianza atribuible a errores de medición se origina en las fluctuaciones azarosas del comportamiento de la persona examinada, que aumentan o disminuyen su puntaje. En este sentido, las distracciones momentáneas, las preocupaciones de índole personal y otros acontecimientos semejantes pueden afectar la estabilidad de los puntajes de tests. Aunque tales influencias fortuitas tienden a atenuarse en una prueba extensa y bien estandarizada, las variaciones en la atención o motivación de las personas al realizar el test siempre pueden afectar su confiabilidad.

La confiabilidad también varía en función de la muestra utilizada para estimarla, lo cual constituye una de las limitaciones más serias del modelo clásico de los tests (Muñiz, 2001). La confiabilidad aumenta al incrementarse la variabilidad de las respuestas y, por consiguiente, se recomienda que las muestras empleadas para verificar la confiabilidad de un test sean lo más heterogéneas posible en aquellas características que generan diferencias entre las personas que la integran (sexo, nivel socioeconómico, nivel educativo, por ejemplo), a los fines de incrementar la variabilidad de las respuestas. En general, se recomienda utilizar muestras grandes y lo más representativas posibles de la población meta del test en cuestión, con un tamaño mínimo de 100 personas (Kline, 2000).

3.4. Dimensiones de la confiabilidad

El concepto de confiabilidad comprende tres dimensiones, cada una de las cuales se relaciona con las diferentes fuentes de error de medición y con distintos métodos para identificarlas.

Cualquier factor no sistemático que incida en el puntaje de un individuo y que no esté relacionado con el constructo que el instrumento intenta medir representa una fuente de error; de manera que debería haber tantas dimensiones de la confiabilidad como condiciones que afectasen las puntuaciones de los mismos (Anastasi y Urbina, 1998). Sin embargo, en la práctica, interesan particularmente sólo tres de estas dimensiones: a) estabilidad; b) consistencia interna y c) confiabilidad inter-examinadores.

a) Si se pretende evaluar en qué grado el puntaje de un individuo en un test está libre de errores de medición causados por cambios personales aleatorios en el examinado (nivel de motivación, por ejemplo), y en algunos casos a los cambios en las condiciones de administración, se hace referencia a la estabilidad de las puntuaciones. Esta dimensión de la confiabilidad está íntimamente relacionada con las características de la variable que se desea medir, puesto que si se están evaluando rasgos que teóricamente tienen cierta estabilidad (por ejemplo, rasgos de personalidad o aptitudes cognitivas), es esperable que las puntuaciones obtenidas en los instrumentos de medición sean también relativamente estables. Si, en cambio, se evalúan estados de ánimo o tiempos de reacción, no resulta relevante atender a la estabilidad temporal de la prueba, ya que teóricamente se espera una modificación de los resultados al aplicarla en distintas ocasiones. Los procedimientos indicados para evaluar la estabilidad temporal de una prueba son el método test-retest y el método de formas equivalentes, cuando ambas formas del test son aplicadas con un intervalo de tiempo (APA, 1999).
b) Si se intenta conocer en que medida la elección de la muestra de ítems que componen la prueba resulta una fuente de error en la medición, se hace referencia a la consistencia interna. Esta dimensión de la confiabilidad alude al grado en que distintas partes o ítems del test miden el mismo constructo o dominio. Los procedimientos para evaluar la consistencia interna de un test son: el método de formas equivalentes, el método de partición en mitades y el método del coeficiente alfa de Cronbach.
c) Si se desea estimar en qué grado la medición de un rasgo a través de un instrumento es independiente de la subjetividad del evaluador se hace referencia a la confiabilidad inter-examinadores. Este tipo de confiabilidad refiere a la objetividad de los datos proporcionados por un test, vale decir, que los individuos obtengan puntuaciones idénticas en sus ejecuciones independientemente de quién sea su examinador. Tanto la calificación de la respuesta de un individuo al test, como la codificación e interpretación deben partir de normas claras y precisas que permitan disminuir el componente subjetivo presente en toda evaluación. El método adecuado para verificar la confiabilidad ínter-examinadores es el acuerdo entre jueces.

3.5. Métodos para verificar la confiabilidad

En los métodos utilizados para corroborar la confiabilidad de un instrumento de medición se pueden distinguir dos instan-

cias. Por una parte es necesario administrar el instrumento a una muestra según un diseño de investigación específico y, por otro lado, los datos que resulten de tal aplicación deben ser analizados mediante procedimientos apropiados para obtener un estadístico que represente la confiabilidad de las puntuaciones del test (Cortada de Kohan, 1999).

a) Test-retest:

Este método consiste en administrar un test en dos oportunidades a la misma muestra de sujetos, con un determinado intervalo entre las dos administraciones, y calcular la correlación entre los puntajes obtenidos en la primera y segunda vez.

Como se explicó previamente, todos los métodos utilizados para valorar la confiabilidad de un test se interesan por el grado de consistencia entre dos conjuntos de puntuaciones obtenidas independientemente, y por lo tanto pueden ser expresados en función de un coeficiente de correlación que exprese el grado de correspondencia o relación entre dos conjuntos de puntuaciones (Anastasi y Urbina, 1998).

El coeficiente de correlación se expresa en un valor que varía entre -1 y 1, donde 0 representa la ausencia total de correlación entre los puntajes, 1 la correlación positiva perfecta (cuando una variable aumenta la otra también lo hace en forma proporcional) y -1 la correlación negativa perfecta (cuando una variable aumenta la otra disminuye en forma proporcional). El coeficiente más comúnmente utilizado es el de correlación momento-producto de Pearson, pero la elección del coeficiente a utilizar depende del nivel de medición empleado por una prueba (nominal, ordinal o intervalar, por ejemplo).

Una de las fórmulas para estimar el coeficiente de Pearson (r) es:

$$r = \frac{N\sum xy \pm (\sum x).(\sum y)}{\sqrt{\left[N.\sum x^2 - (\sum x)^2\right]\left[N.\sum y^2 - (\sum y)^2\right]}}$$

Donde X e Y representan los pares ordenados de puntuaciones correspondientes a la variable X y a la variable Y, y N es el número total de casos examinados. En un estudio de estabilidad, por ejemplo, X equivale a la puntuación de un individuo en la primera administración, Y a la puntuación del mismo individuo en la segunda administración y N el número total de personas evaluadas en ambas administraciones. Como puede verse en la fórmula, para el cálculo de este coeficiente trabajamos con la sumatoria ($\sum$) de estas puntuaciones, no con los puntajes individuales.

Cuando un coeficiente de correlación es utilizado para estimar la estabilidad de las puntuaciones de un test, también suele denominarse coeficiente de estabilidad.

A modo de ejemplo del método test-retest, describimos el estudio de estabilidad realizado en la Escala de Inteligencia para Niños de Wechsler. Se administró el test a una muestra de 353 niños de los Estados Unidos, de 6 a 15 años de edad, estando representados ambos sexos y diferentes etnias. Se dejó transcurrir un tiempo que osciló entre los 12 y los 63 días (con un intervalo mediano de 23 días) y se aplicó nuevamente la prueba. Se correlacionaron los puntajes con el coeficiente de Pearson obteniéndose los siguientes resultados:

Tabla 3.1. Resultados de un estudio test-retest en el WISC III

Grupos de edad	Escala completa	Subtests de ejecución	Subtests verbales
6-7	0, 92	0,86	0,90
10-11	0,95	0,88	0,94
14-15	0,94	0,87	0,94

De acuerdo con estos resultados, el WISC III evidencia una adecuada estabilidad de sus puntuaciones en esta muestra (todos los valores son superiores a 0,80).

Si bien la lógica del método test-retest es sencilla, su aplicación presenta algunos inconvenientes. En primer lugar, casi siempre es incómodo para los examinados someterse a un mis-

mo test en dos oportunidades, lo cual puede generar una disposición desfavorable en la segunda evaluación. Tratar con este problema exige mucha competencia del examinador para generar una adecuada motivación de las personas evaluadas mediante este procedimiento. Si se prevé una devolución de los resultados del test, puede ser útil efectuarla luego de la segunda administración para garantizar un nivel mínimo de motivación en los examinados. Por otro lado, si el intervalo de tiempo transcurrido entre las dos aplicaciones es muy corto, en tests que miden habilidades pueden presentarse problemas relacionados con el efecto de la práctica y la memoria de los sujetos evaluados, obteniéndose una correlación falsamente alta entre las dos aplicaciones. Si, en cambio, el lapso de tiempo es muy prolongado, se corre el riesgo de que las diferencias entre las puntuaciones se deban a cambios reales de los sujetos examinados en la variable que está en estudio, más que a una escasa confiabilidad del test.

El tiempo transcurrido entre una y otra administración debería delimitarse atendiendo a las características de la variable medida y de la población meta del test. Por ejemplo, si el estudio de estabilidad se realiza con niños pequeños, se recomienda que el intervalo de tiempo sea relativamente breve, puesto que en la infancia los cambios progresivos del desarrollo tienen un ritmo más acelerado en la mayoría de las variables psicológicas. El intervalo temporal entre una y otra administración del test no es rígido, y se establece de acuerdo a diversos criterios relacionados con las características del constructo medido y la población meta (Anastasi y Urbina, 1998).

b) Formas equivalentes:

A través de este método, también denominado "formas paralelas", se puede evaluar la consistencia interna pero también la estabilidad temporal de un conjunto de puntuaciones. El procedimiento básico consiste en administrar dos formas equivalentes de un test a un mismo grupo de individuos. En el caso de que este método se utilice para verificar la estabilidad, la administración de la segunda forma se realiza transcurrido un tiempo a partir de la administración de la primera forma, y posteriormente se correlacionan los resultados obtenidos.

Si bien este método es el más completo para evaluar la confiabilidad, puesto que permite controlar la mayor cantidad posible de fuentes de error aleatorio (distinta muestra de ítems, distintas condiciones físicas y mentales de los examinados, diferente situación ambiental, posiblemente distinto administrador), en la práctica presenta algunos inconvenientes.

Para ser consideradas equivalentes, dos pruebas deben reunir ciertos requisitos, tales como tener las mismas características formales (cantidad de ítems, escala de respuesta, etc.) y estadísticas (tener medias y desviaciones estándar semejantes, coeficientes de correlación elevados entre ambas formas, etc.) (APA, 1999).

El procedimiento ha sido simplificado con la informatización de los tests, que posibilita utilizar las formas computarizadas como versiones alternativas de las originales en lápiz y papel. No obstante, tal como lo expresan las normas especiales de la APA para tests computarizados, este tipo de tests no puede ser tratado automáticamente como forma paralela de los tests de lápiz y papel, sino que debe demostrarse el cumplimiento de los requisitos anteriormente mencionados.

Un ejemplo del método de formas paralelas se puede encontrar en el manual del Test de Aptitudes Diferenciales (Bennett, Seashore y Wesman, 2000). A una muestra de 215 adolescentes españoles de ambos sexos se les administró la forma S del Subtest de Velocidad y Precisión y, transcurrido un breve lapso, se les administró la forma T del mismo subtest. La tabla siguiente presenta los resultados de esta investigación, donde se pueden observar los coeficientes de correlación entre ambas formas y las medias y desviaciones estándar de los dos subtests para grupos diferenciados por sexo. Los resultados permiten concluir que la confiabilidad evaluada a través del método de formas equivalentes es satisfactoria para el subtest Velocidad y Precisión. En efecto, como puede observarse en la tabla, ambas formas se correlacionan adecuadamente y las medias y desviaciones estándar son semejantes en ambas aplicaciones del test.

Tabla 3.2. Estudio de Formas Equivalentes en el DAT

			Forma S		Forma T		
Sexo	**N**	**r**	**M**	**s**	**M**	**s**	**Valores t**
Varones	112	0,89	45.2	10,7	46.5	10,8	0,11
Mujeres	103	0,90	50.5	10,3	52.9	10,1	0,20

c) Partición en mitades:

A través de este método se verifica la consistencia interna de las puntuaciones de un test, es decir, el grado en que las diferentes partes del test miden la misma variable. Se administra el test en una ocasión a una muestra de individuos y posteriormente se divide la prueba en dos mitades comparables, obteniendo de esta manera dos puntuaciones para cada individuo de la muestra. Finalmente, se correlacionan las puntuaciones correspondientes a ambas mitades del test por medio de un coeficiente de correlación.

Este método fue popular antes de que se dispusiera de computadoras personales, debido a que los estadísticos requeridos son más fáciles de calcular manualmente que el coeficiente alfa que se presenta en el apartado siguiente. Actualmente, el método de partición en mitades es poco empleado para verificar la consistencia interna de una prueba (Nunnally y Bernstein, 1995).

La dificultad inicial de este procedimiento es lograr que las mitades obtenidas sean realmente comparables. Muchos de los tests son construidos con un nivel de dificultad creciente, y si se divide el test en la primera y segunda mitad seguramente éstas no resultarían comparables. Aun cuando no sean pruebas con dificultad creciente de sus ítems, otros factores pueden obstaculizar el contar con dos mitades estrictamente comparables. En efecto, los examinados probablemente se vean más afectados por la fatiga hacia el final de la prueba, lo cual podría incidir en mayor medida en los puntajes de la segunda mitad.

El criterio habitualmente adoptado para dividir la prueba es el de separar los ítems del test en dos mitades, una de ítems pares y la restante de ítems impares. Un método más riguroso,

pero que requiere mayor esfuerzo del investigador, exige el apareamiento de todos los ítems con un cierto criterio estadístico, asignándolos luego al azar a cada una de las partes o mitades del test.

No obstante, en el método de partición en mitades, el coeficiente de correlación obtenido (*r* de Pearson, por ejemplo) expresa la confiabilidad de una sola de las mitades, por lo que calcular la confiabilidad de la prueba completa requiere el uso de un estadístico adicional, la fórmula de corrección de Spearman-Brown:

$$r_n = \frac{n.r}{1 + (n-1).r}$$

Donde:

r_n = coeficiente de confiabilidad para el test total (confiabilidad ajustada por la fórmula de Spearman-Brown)

r = coeficiente de correlación entre las dos mitades

n = el número de veces que el test se acorta (en el caso de dos mitades n = 2)

Esta fórmula resulta en un sensible aumento del coeficiente de correlación inicial entre las dos mitades, puesto que las puntuaciones del test completo son siempre más confiables que las de sus subdivisiones debido al mayor número de ítems.

d) Métodos de covarianza de los ítems:

Estos métodos comparten con el anterior dos aspectos importantes: por un lado, permiten verificar la consistencia interna de los puntajes del test y, por otra parte, requieren una sola administración de la prueba (Thorndike, 1989).

A partir de una única aplicación del test a una muestra se obtiene una estimación del grado de covarianza de los ítems, utilizando como estadístico el coeficiente alfa de Cronbach o la fórmula alternativa de Kuder Richardson (KR-20), cuando se trabaja con ítems dicotómicos (verdadero-falso, por ejemplo).

La fórmula del coeficiente alfa de Cronbach (1951) es:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\Sigma s^2}{St^2}\right)$$

Donde:

k = número de ítems de la prueba
$\sum s$ = sumatoria de la varianza de cada ítem
St = varianza del total de las puntuaciones del test

El coeficiente alfa puede considerarse como la media de todas las correlaciones de partición por mitades posibles (Cohen y Swerdlik, 2000). Según Muñiz (2001), el coeficiente alfa expresa el grado de covariación de los ítems de un test, o en qué medida los diferentes ítems de un test miden una misma variable.

En la actualidad, el coeficiente alfa es el estadístico más popular para estimar la consistencia interna de una prueba. Es un coeficiente apropiado en tests que contienen ítems multipunto (con varias alternativas de respuesta). Como se indicó más arriba, cuando los ítems son dicotómicos habría que utilizar la fórmula alternativa KR-20

$$KR_{20} = \frac{k}{k-1}\left(1 - \frac{\Sigma_{pq}}{St^2}\right)$$

Donde *k* es el número de reactivos del test, *p* es el porcentaje de casos que acierta el ítem, *q* es igual a 1-p y St^2 es la varianza de las puntuaciones del test.

Debe destacarse que tanto el método de partición en mitades como el coeficiente alfa son inapropiados para corroborar la confiabilidad de tests de velocidad o tiempo limitado (Anastasi y Urbina, 1998). En esos casos deben utilizarse métodos alternativos, como el test-retest o el de formas equivalentes.

A continuación se presenta una tabla que ilustra el modo de presentación de los resultados de un análisis de consistencia interna utilizando el coeficiente alfa en el Inventario de Autoeficacia para Inteligencias Múltiples (Pérez, 2001). El instrumento fue aplicado a una muestra de estudiantes secundarios de ambos sexos (N = 917) y se obtuvieron los siguientes coeficientes:

Tabla 3.3. Coeficiente alfa para cada escala del IAMI

Escalas	**a**
Intrapersonal	0,86
Naturalista	0,91
Lingüística	0,86
Matemática	0,89
Espacial	0,91
Cinestésica	0,93
Musical	0,93
Interpersonal	0,85

Como puede apreciarse, los resultados indican un nivel de consistencia interna adecuado para todas las escalas (todos los a son superiores a 0,80).

Cuando los ítems de un test o escala son numerosos (superiores a 30) el coeficiente alfa tiende a ser demasiado elevado (Cortina, 1993). En ese caso se recomienda el uso adicional del coeficiente de correlación inter-ítem, menos influido por el número de ítems de una escala. La magnitud recomendable del coeficiente de correlación ínter-ítem debe situarse entre 0,15 y 0,50 (Carretero-Dios y Pérez, 2005).

e) Acuerdo entre examinadores:

La dimensión evaluada por este método es la confiabilidad entre examinadores. El método consiste en administrar un test a una muestra, entregar los resultados (protocolos de respuesta) del test a un conjunto de jueces que los puntuarán independientemente. A continuación, se verifica el grado de acuerdo que alcanzan los jueces luego de leer, registrar y codificar los mismos datos (Murat, 1985). Naturalmente, este procedimiento

no se aplica en tests que se puntúan de manera objetiva (de opción múltiple, por ejemplo) y sólo adquiere importancia cuando interviene el criterio del examinador en el proceso de calificación.

El coeficiente utilizado para estimar el grado de acuerdo entre jueces estará determinado por el tipo de escala que ellos empleen para calificar las respuestas de los sujetos a los ítems de la prueba. Los coeficientes comúnmente utilizados son el índice kappa, cuando se trata de escalas nominales, y en el caso de escalas ordinales o intervalares, los estadísticos kappa modificado, *w* de Kendall, o el coeficiente de correlación intraclase.

El coeficiente kappa nos permite estimar concordancia entre observadores, es decir, hasta qué punto los jueces coinciden en su puntuación (Muñiz, 2001) considerando el porcentaje de acuerdos que se observarían solamente por azar. La fórmula de kappa es:

$$K = \frac{Fc - Fa}{N - Fa}$$

En la fórmula precedente, Fc son las frecuencias de coincidencias o número de casos en los que las clasificaciones de ambos jueces coinciden. Se obtiene sumando las celdas que representan los casos que fueron evaluados de la misma manera por ambos jueces. Fa son las frecuencias de azar, o número de casos en que cabe esperar que las clasificaciones de los jueces coincidan por mero azar, y se obtienen mediante la sumatoria de los productos de los subtotales de cada categoría sobre el número de casos. N es el número total de casos evaluados por los jueces.

Para el caso de dos jueces que calificasen, por ejemplo, un ítem de un test de vocabulario (1 punto por definición correcta de la palabra o sinónimo apropiado, 0 punto para las respuestas erróneas) en 200 protocolos, tendríamos una matriz como la que sigue a continuación.

En la tabla 3.4. se indican los acuerdos y desacuerdos entre los jueces. Así, por ejemplo, se observa que el Juez A consideró 94 protocolos como correctos y 106 como incorrectos, mientras que el Juez B estimó que 98 eran correctos y 102 erróneos. Por

Tabla 3.4. Puntuación de un ítem por dos jueces en forma independiente

	Puntuación del Juez A		
Puntuación del juez B	**Respuesta correcta**	**Respuesta incorrecta**	**Total**
Respuesta Correcta	85	13	**98**
Respuesta Incorrecta	9	93	**102**
Total	**94**	**106**	**200**

otra parte, el juez A evaluó 9 casos como correctos que el juez B consideró incorrectos, mientras que en los restantes 85 acordaron. Del mismo modo, ambos jueces acordaron en su estimación de 93 protocolos incorrectos. Para este último caso el desacuerdo estuvo dado por los 13 protocolos que el juez A evaluó como incorrectos y el juez B como correctos.

Aplicando a la tabla precedente la fórmula de kappa se verifica que:

Fc = 85 + 93= 178
Fa = 94.98/200 + 106.102/200= 100, 12
K = 178 – 100,12/ 200 – 100, 12= 0,78

La máxima concordancia posible corresponde a $k = 1$. El valor $k = 0$ se obtiene cuando la concordancia observada se corresponde con la esperada exclusivamente del azar. A la hora de interpretar el valor de *k* es útil disponer de una escala como la siguiente aunque, en general, se considera adecuado un coeficiente de acuerdo de 0,80 o superior, lo que sugeriría que el test en cuestión permite una interpretación unívoca de sus resultados independientemente del evaluador.

Es importante señalar que la fórmula original de *k* se aplica solamente en el caso de que intervengan dos jueces, aunque existen alternativas a esta fórmula de cálculo, utilizable cuando intervienen más jueces, tales como la fórmula modificada de *k* propuesta por Fleiss (1971) o el coeficiente de correlación intraclase (Hogan, 2004).

Tabla 3.5. Interpretación del coeficiente kappa

Valor de k	**Fuerza de la concordancia**
< 0,20	Pobre
0,21 – 0,40	Débil
0,41 – 0,60	Moderada
0,61 – 0,80	Buena
0,81 – 1,00	Muy buena

La confiabilidad inter-examinadores puede evaluarse por ítem, y en ese caso estaríamos verificando el grado de acuerdo de los jueces para puntuar un ítem de un test determinado (con 2, 1 ó 0, por ejemplo). También se puede obtener una estimación de la confiabilidad inter-examinadores para la puntuación total asignada por los jueces a una escala donde todos los ítems se califican con un componente de subjetividad.

A continuación se ejemplifica el caso de un estudio de concordancia para más de dos jueces. Tres psicólogos clínicos podrían calificar a 50 pacientes utilizando una escala de desadaptación con una puntuación que va de 0 (sin desadaptación observable) a 20 (gravemente desadaptado). En este caso tendríamos una tabla como la siguiente:

Tabla 3.6. Datos para el estudio de la confiabilidad entre tres examinadores

	Jueces		
	A	**B**	**C**
Paciente			
1	14	8	13
2	8	7	7
3	16	15	14
-	-	-	-
50	5	4	6

Para expresar el grado de acuerdo entre los tres jueces del ejemplo, podrían estimarse las correlaciones entre todas las posibles combinaciones de calificadores (A *vs*. B; A *vs*. C; B *vs*. C) y luego promediar esas correlaciones. Sin embargo, esto no sería adecuado puesto que las series de calificaciones de los jueces pueden estar fuertemente correlacionadas pero ser poco concordantes. Una estimación más directa y apropiada del acuerdo entre examinadores es suministrada por el coeficiente de correlación intraclase (CCI) anteriormente mencionado, que se calcula a partir de los estadísticos generados por el análisis de varianza, con varias fórmulas alternativas (Hogan, 2004). El CCI también puede calcularse directamente en SPSS desde el menú de confiabilidad.

Tabla 3.7. Resumen de métodos y estadísticos para evaluar la confiabilidad

Dimensión de la confiabilidad	**Método**	**Nº de sesiones de administración de la prueba**	**Estadístico**
Estabilidad	Test-retest	2	r de Pearson
	Formas paralelas	2	r de Pearson
Consistencia interna	Formas paralelas	1	r de Pearson
	Partición en mitades	1	r y fórmula de corrección Spearman-Brown
	Coeficiente alfa	1	Alfa, Kuder-Richardson
Confiabilidad entre examinadores	Acuerdo entre examinadores	1	Kappa, w de Kendall, coeficiente de correlación intraclase

La tabla anterior resume los métodos y estadísticos más utilizados para verificar la confiabilidad de las puntuaciones de tests, considerando sus diversas dimensiones.

Un aspecto importante a considerar en un estudio de confiabilidad es el tamaño del coeficiente de confiabilidad. Cuando un test va a utilizarse sólo con fines de investigación, un coeficiente de 0,70 es suficiente. Si los tests se emplean para la toma de

decisiones que afectan a las personas examinadas (contextos de clasificación, por ejemplo) son deseables valores de 0,90 o superiores. En general, los resultados de tests de ejecución máxima son más confiables que aquéllos provenientes de tests de respuesta típica, pero siempre son recomendables valores de 0,80 o superiores para tests utilizados en contextos aplicados (Hogan, 2004).

3.6. Confiabilidad y puntuaciones individuales

Recordemos una vez más que la puntuación obtenida por un individuo en cualquier test está compuesta por la puntuación verdadera más el error de medición. Es decir, el puntaje verdadero es el puntaje que realmente habría obtenido en un test una persona determinada si hubiéramos podido eliminar todos los factores de error.

Debido a que la puntuación verdadera es una puntuación teórica, es imposible conocer su valor directamente y sólo podemos estimar (en términos probabilísticos) la ubicación del puntaje verdadero con un cierto grado de confianza. Para estimar su ubicación probable partimos del supuesto de que si el sujeto realizara una cantidad infinita de pruebas equivalentes, las puntuaciones en esos tests tenderían a distribuirse de manera normal con la puntuación verdadera del individuo como la media de esa distribución. En ese contexto, la desviación estándar de esa distribución se denomina error estándar de medición (*EEM*). Este estadístico se calcula a partir del coeficiente de confiabilidad mediante la siguiente fórmula:

$$EEEM = s_t.\sqrt{1 - r_{xx}}$$

Donde:

s = desviación estándar de las puntuaciones del test en la muestra de estandarización

r_{xx} = coeficiente de correlación test-retest

Si una escala de un inventario de intereses vocacionales (Escala Realista del Self-Directed Search, por ejemplo) tiene un coeficiente de estabilidad de 0,84 y una desviación estándar de 10, entonces:

$$10.\sqrt{1 - 0{,}84} = 4$$

El error estándar de medición es fundamental para poder estimar la ubicación aproximada de la puntuación verdadera y permite interpretar las puntuaciones individuales de un test. Esto se logra a partir del establecimiento de intervalos de confianza, vale decir, un rango de puntuaciones que probablemente contengan la puntuación verdadera. Debido a que se asume una distribución normal cuya media está representada por la puntuación verdadera, podemos suponer que si repitiéramos el test muchas veces, el 68% de los puntajes que obtenga esta persona va a estar entre ± 1 EEM de la puntuación verdadera. Así, por ejemplo, si el puntaje observado de un estudiante es 40 en un test (X = 40) y el error estándar de medición es 4 (EEM = 4), si repitiéramos el test muchas veces, aproximadamente el 68% de los puntajes (68,26%) que obtenga esta persona va a estar entre 36 y 44, es decir:

X ± 1. EEM
40 + 1, 4

Si repitiéramos el test muchas veces, el 95% de los puntajes va a estar entre ± 1,96 EEM de la puntuación verdadera. De esta forma, podemos tener una probabilidad del 95% de que el puntaje verdadero del individuo va a estar comprendido entre 32 y 48 (valores redondeados), es decir 1,96 EEM por arriba o por debajo del valor obtenido en el test (Anastasi y Urbina, 1998), o sea:

X ± 1,96. EEM
40 + 1,96. 4

Como puede inferirse de su formula de cálculo, el error estándar de medición está en relación inversa con el coeficiente de confiabilidad; mientras mayor sea la confiabilidad, menor será el error estándar de medición y más confianza se puede tener en la exactitud o precisión del puntaje observado, lo cual se ve reflejado en la menor amplitud de los intervalos de confianza estimados. Para entender mejor este punto, considérese el caso anterior en el cual se obtuvo un coeficiente de confiabilidad de 0,84 y un error estándar de 4. Suponiendo que una persona obtuvo en dicha prueba una puntuación de 70, el usuario del test puede estar seguro en un 95% de que el puntaje verdadero de esta persona cae en el rango de 62 a 78. Esto se debe a que el intervalo de confianza de 95% se establece a partir de la puntuación observada de 70 más o menos 1,96 multiplicado por el error estándar de medición.

$$70 \pm 1{,}96.\ 4 = 70 \pm 8$$

Si el coeficiente de confiabilidad hubiese sido de 0,90, el error estándar hubiese sido 3. Esto implica que para la misma puntuación (x = 70) el administrador del test puede confiar con una probabilidad del 95% de que el puntaje verdadero de esta persona estaría incluido en el intervalo de 64 a 76. Como puede verse en el ejemplo, un mayor coeficiente de confiabilidad trae aparejada una disminución en el error estándar y una menor amplitud en los intervalos de confianza, lo cual implica en definitiva una mayor precisión en la medición.

Los intervalos de confianza resultan de utilidad en los tests empleados en selección de personal, cuando debe decidirse la situación de los individuos cuyos puntajes están cercanos al punto de corte (por encima o por debajo de él). Si se desea comparar puntuaciones, ya sea de un mismo individuo en dos ocasiones o de dos individuos entre sí, es necesario emplear el error estándar de la diferencia, una medida estadística útil para determinar qué tan grande debe ser una diferencia en puntuaciones para ser considerada estadísticamente significativa (Cohen y Swerdlik, 2000).

Comunicar solamente el coeficiente de confiabilidad sin referencia a estimaciones del error de las mediciones no aporta demasiada información al usuario de un test. Tal como expresa la norma 2.1 de la APA (1999), para cada puntuación total, parcial o combinación de puntuaciones que será interpretada, debe informarse acerca de los coeficientes de confiabilidad y el error estándar de medición.

El error estándar de medición se conceptualiza de manera diferente en el contexto de la teoría de repuesta al ítem, adquiriendo un carácter condicional. Esto significa que no es el mismo para todas las personas evaluadas por un test, sino que dependerá del nivel en el atributo del individuo evaluado. Este tema se trata en el último apartado de este capítulo.

3.7. Confiabilidad en la teoría de respuesta al ítem (TRI) y en los tests con referencia a criterio (TRC)

Las pruebas construidas en base a la TRI también deben asegurar la confiabilidad de sus puntuaciones. No obstante, existen algunas diferencias fundamentales en los procedimientos para estimar la confiabilidad en este tipo de pruebas. En el último capítulo se revisan los conceptos fundamentales de la TRI, y es conveniente repasar este apartado luego de haber leído ese capítulo. Por ahora sólo unos conceptos preliminares.

El análisis de la consistencia interna desde la TCT se basa en el análisis del funcionamiento de los ítems dentro de las pruebas, asumiendo que un coeficiente de confiabilidad es el mismo para todas las personas a las que se aplica la prueba. No obstante, en la TRI los ítems operan en forma independiente (Hogan, 2004), y se asume que la precisión de una prueba no es la misma para todas las personas, ya que la precisión es una función del nivel de la persona en la variable medida.

Por este motivo, la estimación de la confiabilidad en el contexto de la TRI se realiza mediante la denominada función de información. La función de información de un ítem indica la precisión de un ítem para medir el rasgo en diferentes niveles del continuo de ese rasgo, mientras que la función de información del test sugiere la confiabilidad de todos los ítems en conjunto para medir el rasgo en diferentes niveles (Hogan, 2004; Baker, 2001). La función de información del test se estima por

la sumatoria de la función de información de todos los ítems del test en todos los niveles del rasgo.

De esta manera, un mismo test puede ser más fiable para algunas personas que para otras, en función de la dificultad del mismo y su adecuación para el nivel de una persona determinada en el dominio o rasgo medido. El error estándar de medición, en el contexto de la TRI, se determina también específicamente para cada nivel de puntuación, por lo cual éste puede diferir entre niveles altos y bajos del rasgo (Hogan, 2004). Este índice se calcula a partir de la función de información.

Más adelante introducimos al lector en la problemática de los tests referidos a criterio (TRC), esto es, las pruebas de rendimiento en un dominio específico de conocimiento (Lengua o Matemática, por ejemplo). En ese caso no se interpretan las puntuaciones con relación a una muestra de referencia (como en las referidas a normas o baremos) sino que se identifica la posición absoluta del sujeto con respecto a algún dominio de conductas previamente definido. En estos tests interesa fundamentalmente comprobar la confiabilidad de las clasificaciones establecidas mediante su utilización respecto a la maestría de dominio de los individuos que están aprendiendo un dominio (expertos vs. no expertos).

En los TRC se establece un punto de corte (puntuación equivalente a la maestría de dominio) en las puntuaciones obtenidas mediante el cual se identifica a los individuos que alcanzan o no los objetivos planteados en un dominio educativo (utilizar eficientemente un procesador de texto, por ejemplo). Para esta finalidad son indicados los coeficientes de acuerdo entre examinadores, tales como el kappa (k) mencionado anteriormente. Cohen (1960) presentó el coeficiente k para expresar el acuerdo en las proporciones de clasificación de los individuos (expertos y no expertos en el dominio) entre dos aplicaciones del test. Este coeficiente sería equivalente al test-retest, con valores comprendidos entre 1 (experto en el dominio) y 0 (no experto). En este caso se aplica la fórmula siguiente, semejante a la explicada previamente:

$$K = \frac{(P \pm P_c)}{(1 \pm P_c)}$$

Donde P es igual a la proporción de personas clasificadas en la misma categoría en las administraciones del test, es decir: (proporción de expertos en las dos administraciones) + (proporción de no expertos en las dos ocasiones) / N; y Pc es la proporción de individuos clasificados que se esperaría por azar (Hogan, 2004). Al igual que en la fórmula de kappa presentada en el apartado de confiabilidad inter-examinadores, el azar está determinado por los totales marginales de la tabla utilizada.

Hay varios índices de confiabilidad apropiados para los TRC cuyo tratamiento excedería el marco de un texto introductorio. Para un examen del tema en español muy completo y autorizado, puede consultarse el texto *Psicometría,* de Martínez Arias (1995).

4
VALIDEZ

Edgardo Pérez - Fabián Olaz

4.1. Introducción

La validez es un aspecto esencial de la medición psicológica y se relaciona con la investigación del significado teórico de las puntuaciones obtenidas por medio de un test (Oliden, 2003).

> Las puntuaciones evidencian propiedades de validez cuando se verifica que el test realmente mide el constructo que pretende medir, justificando adecuadamente las inferencias realizadas en función de sus resultados (Nunnally, 1991).

Recordemos que "constructo" es la representación abstracta de un conjunto de comportamientos relacionados. Así, por ejemplo, "depresión" o "aptitud matemática" son constructos. Se podría construir un inventario de 20 ítems para medir depresión, pero alguna dimensión del constructo (como el componente emocional de la depresión) podría estar escasamente representada por el contenido de ese test. Esa dimensión no está adecuadamente representada por los ítems de un test ocasiona la *sub-representación del constructo.* Por otra parte, es probable que las puntuaciones en un inventario reflejen una tendencia a dar respuestas socialmente deseables (aceptables) por parte del examinado. Este aspecto de las puntuaciones no relacionadas con el verdadero propósito de medición del test se denomina *varianza irrelevante del constructo* (Hogan, 2004).

La situación ideal en lo referente a la validez es que un test represente adecuadamente y mida la varianza relevante del

constructo, o expresado de otro modo, que las interpretaciones de los resultados de una prueba estén libres de sesgo de medición. Según Oliden (2003), la teoría de la validez se relaciona con el concepto de sesgo, definido como un error sistemático que produce distorsión en las puntuaciones adulterando su significado teórico. Tal como afirma Muñiz (1998), el hecho que las puntuaciones de un test sean confiables es una condición necesaria pero no suficiente para que sean válidas.

A pesar de su importancia, el concepto de validez es uno de los más complejos y controvertidos de la teoría de los tests (Angoff, 1988; APA, 1999). Una breve introducción histórica puede contribuir a esclarecer la significación actual de este concepto. Para Oliden (2003), la evolución de las teorizaciones acerca de la validez, desde un enfoque operacional y pragmático hasta la concepción contemporánea, refleja los cambios que se fueron dando en la ciencia psicológica en general.

En la historia del concepto de validez pueden identificarse tres etapas principales. Una primera etapa, operacional, en la que predomina una perspectiva exclusivamente pragmática de las aplicaciones de los tests. Este enfoque coincide con el operacionalismo dominante en la epistemología de la primera mitad del siglo pasado y se manifiesta en la noción de validez como sinónimo de la correlación entre las puntuaciones de un test y algún criterio que el test intenta predecir (Martínez Arias, 1995). Como señala Angoff (1988), la concepción de validez con un sentido meramente predictivo dominó el escenario de la psicometría hasta los años cincuenta.

Posteriormente se comprendió que este concepto de validez exclusivamente ligado a la predicción de criterios externos no era útil para muchos tests en los que ellos mismos constituyen su propio criterio (por ejemplo, en pruebas de rendimiento) y esto condujo a introducir el concepto de validez de contenido. Otro cambio importante se produjo con la aparición del clásico artículo de Cronbach y Meehl (1955) donde se presentó por primera vez el concepto de validez de constructo y se caracterizó a esta última como el aspecto fundamental e inclusivo de las restantes dimensiones de la validez (Martínez Arias, 1995). Esta publicación inició un segundo estadio teórico, en el cual asume un papel fundamental la teoría psicológica. En esta fase se diferencian tres tipos de validez: de constructo, de contenido y predictiva.

Por último, el período actual o contextual se caracteriza por una extensión de la concepción anterior, a la que se agrega la importancia otorgada al uso propuesto para el instrumento. Esto significa que, en realidad, nunca se valida un test en sí mismo sino que su validez se verifica para determinados propósitos. En esta nueva perspectiva ya no se habla de distintos tipos de validez sino de un proceso de recolección de diferentes tipos de evidencia para un concepto unitario.

Esta concepción contemporánea de validez se refleja en la última versión de las Normas Técnicas para los Tests Psicológicos y Educativos (APA, 1999), donde se la define como la adecuación, significación y utilidad de las inferencias específicas hechas a partir de las puntuaciones de los tests. Como expresamos, la validez es un concepto unitario y siempre se refiere al grado en que la evidencia empírica apoya las inferencias realizadas en función de los resultados de un test. La APA (1999) propuso cinco tipos de evidencia de validez, basadas en: el contenido del test, la estructura interna del test, el proceso de respuestas al test, las relaciones con otras variables externas al test y las consecuencias de su aplicación.

De esta manera, para verificar la validez de las inferencias realizadas a partir de las puntuaciones de un test se utilizan procedimientos semejantes a los implementados para contrastar cualquier hipótesis científica, vale decir la recolección de evidencias que confirmen o refuten esas inferencias.

> El producto final del proceso de validación es la medición de un constructo que: a) esté bien definido en términos de una variedad de observaciones y, eventualmente, b) se correlacione con otros constructos de interés.

A continuación nos ocupamos más detenidamente de las fuentes de evidencia de validez de los tests.

4.2. Fuentes de evidencia de validez

Fuentes internas de evidencia

Según Oliden (2003), existen fuentes de evidencia internas y externas. La primera categoría se relaciona con el test y sus componentes (ítems) en sí mismos. La lógica implícita de las evidencias incluidas en esta categoría se relaciona con el primer objetivo del proceso de validación de las puntuaciones de un test, es decir, medir un constructo con un significado unívoco, estrictamente definido.

a) Evidencia basada en el contenido del test

Para Murat (1985), este tipo de evidencia se obtiene demostrando que el contenido (ítems) del test es una muestra representativa del constructo o dominio respecto del cual se desea hacer alguna inferencia. Debe existir correspondencia entre el contenido del test y el dominio (área de comportamiento o conocimiento) que éste pretende medir (Hogan, 2004). El razonamiento implícito es que, si los ítems de un test son representativos de un dominio particular, el desempeño del sujeto en el mismo puede generalizarse a todo el dominio (Herrera Rojas, 1998).

Esta evidencia es más factible de ser obtenida en las pruebas de rendimiento (tests referidos a criterio, por ejemplo), donde se necesita verificar la representatividad y relevancia del contenido del test con respecto a los objetivos, actividades, conocimientos y destrezas que se proponen en un curso o ciclo de enseñanza (la ortografía de palabras habituales en la escuela primaria por ejemplo). Es evidente que se requiere un plan detallado y extenso que abarque el contenido y los objetivos del curso en cuestión. La coincidencia de un plan detallado de prueba con los objetivos académicos se puede apreciar sólo cuando esos objetivos están definidos clara y explícitamente (Thorndike, 1989). En el capítulo 6 sobre construcción de tests analizaremos más detenidamente el proceso de definición del contenido de un test en pruebas de rendimiento.

Los tests que evalúan rasgos latentes (intereses o aptitudes por ejemplo) poseen menos representatividad en relación con el dominio de comportamiento que intentan medir, puesto que no se basan (como los tests de rendimiento) en un dominio de conocimiento específico (el programa de una asignatura, por ejemplo). En efecto, en este último caso es más sencillo realizar un muestreo de objetivos, contenidos y actividades y conseguir que el contenido del test sea representativo del dominio de interés. Aunque en las fases iniciales de elaboración los constructores de tests de rasgos latentes también evalúan la adecuación y congruencia del contenido, la validación final de los tests que miden constructos se relaciona más con los otros tipos de evidencia (Anastasi y Urbina, 1998). No obstante, el primer paso para los constructores de cualquier tipo de test incluye especificaciones adecuadas del dominio de contenido que el test intenta representar en función de sus usos propuestos.

De acuerdo con la APA (1999), los métodos para reunir evidencia de contenido se apoyan mayoritariamente en el juicio de expertos, que permite confirmar la relación entre los ítems del test y el dominio o constructo a medir, pero también pueden emplearse otros procedimientos lógicos y empíricos facilitados por la tecnología computacional, lo que permite generar ítems que difieran sistemáticamente en varias pautas del dominio.

Según Martínez Arias (1995), en la validación relacionada con contenido deben realizarse las siguientes operaciones:

a. Definición del dominio de conocimiento o comportamiento a medir.
b. Identificación de expertos en ese dominio.
c. Juicio de los expertos acerca del grado en que el contenido del test es relevante y representativo del dominio.
d. Procedimiento estadístico para resumir los datos de la fase precedente.

Cuando se entregan a los jueces los ítems preliminares de un test es conveniente adjuntar una forma estandarizada de calificación. Los jueces revisan de manera independiente cada uno de los ítems, utilizando esa guía preestablecida (Herrera Rojas, 1998).

Tabla 4.1. Ejemplo de guía para revisión de ítems por parte de expertos

Ítem	Pertinencia	Relevancia	Aspectos formales (sintácticos, p. e.)	Observaciones
5	SÍ	NO	Adecuado	El ítem es congruente con el contenido de la escala, pero mide aspectos secundarios del atributo.
7	SÍ	SÍ	Inadecuado	El ítem es pertinente y relevante, pero inadecuado para el nivel de maduración de los sujetos.

Se han realizado algunos intentos para desarrollar índices objetivos (equivalentes al coeficiente de correlación, por ejemplo) que permitan verificar la evidencia de contenido. Hambleton (1994) propuso suministrar a los expertos una lista de las dimensiones medidas por el test y presentarles cada ítem en una ficha separada. De este modo, cada experto compara los ítems con la lista numerada de dimensiones del constructo a medir e indica al lado de cada ítem el número de dimensión correspondiente según su opinión. Un coeficiente de concordancia, tal como el kappa (examinado en el capítulo anterior), permitiría evaluar el grado de acuerdo entre los jueces en relación con cada ítem.

Así, por ejemplo, en un test construido para medir autoeficacia para la enseñanza, los investigadores delimitaron una serie de dimensiones, tales como: autoeficacia para motivar a los alumnos, para comprometer a los padres en el proceso de enseñanza y para el uso de recursos didácticos, entre otros. Se suministró a los jueces el listado de ítems (por ejemplo "comienzo mis clases estableciendo relaciones entre los contenidos de la asignatura y algunos sucesos de la vida cotidiana de los estudiantes") y la definición de cada dimensión; posteriormente, los jueces debían indicar a qué dimensión correspondía cada uno de los ítems según su criterio.

Los expertos también pueden juzgar la calidad formal de los ítems de un test (corrección gramatical, claridad, adecuación del contenido a la población meta) utilizando una escala numérica de 1 a 5, por ejemplo. En este caso, se deberían retener aquellos ítems con valores promedio (media o mediana) más altos. Los ítems con puntuaciones más bajas deben ser descartados o modificados en su redacción.

Como puede apreciarse en la tabla precedente, también es recomendable solicitar a los jueces observaciones complementarias acerca de la calidad de los ítems (columna de observaciones), las cuales seguramente serán útiles para modificar el fraseo de algunos ítems antes de su redacción final.

No debe confiarse exclusivamente en el juicio de expertos para evaluar la calidad y pertinencia de los ítems. Un estudio piloto con una muestra pequeña de características semejantes a la población meta del test seguramente también proporcionará información útil para la redacción definitiva de los ítems (por ejemplo, si hay instrucciones poco claras, alguna palabra de difícil comprensión).

b) Evidencia basada en la estructura interna del test

Las evidencias relacionadas con la estructura interna de un test indican si las relaciones entre los ítems y las dimensiones (factores, escalas) permiten confirmar la existencia de los constructos que el test pretende medir. El marco conceptual de un test puede proponer una dimensión unitaria de comportamiento o varios factores. Una encuesta podría construirse para medir salud orgánica y emocional, por ejemplo; si las intercorrelaciones entre los ítems confirman esos dos factores teóricos propuestos, ésta es una información relevante para la evidencia de validez relacionada con la estructura interna del test (APA, 1999).

Es necesario verificar que estadísticamente los ítems se agrupen del modo que se predice teóricamente, y para esa finalidad el procedimiento ideal es el análisis factorial (Carretero-Dios y Pérez, 2005). En efecto, el análisis factorial fue desarrollado para identificar constructos psicológicos y es especialmente relevante para obtener evidencia de la estructura interna de un test.

El análisis factorial es un método estadístico para analizar las intercorrelaciones entre datos observables (Martínez Arias, 1995). Por ejemplo, si se aplican 20 ítems a 200 personas, el primer paso consiste en estimar las correlaciones de cada ítem con los demás. Al observar la matriz de correlaciones obtenidas se demostrarán ciertas agrupaciones entre los ítems, lo que permitirá inferir la existencia de rasgos (factores) comunes.

De este modo, si en un test de aptitudes cognitivas los ítems que evalúan vocabulario presentan correlaciones altas entre sí y bajas correlaciones con los otros agrupamientos de ítems (factores), podemos inferir la existencia de un factor de razonamiento verbal. En el análisis factorial, esencialmente, se reduce el número de variables inicialmente contempladas y el comportamiento de cada individuo puede describirse con referencia a un número relativamente pequeño de factores o rasgos comunes (Anastasi y Urbina, 1998).

Algunas de las principales teorías relacionadas con constructos psicológicos, tales como el modelo de los cinco factores de la personalidad y la teoría de los tres estratos de inteligencia, son producto de análisis factoriales que, en el primer caso, permitieron reducir un enorme número de adjetivos descriptivos de rasgos y, en el segundo, dotar de estructura a más de 60 habilidades específicas. En el capítulo relacionado con construcción de tests examinaremos con mayor detalle el análisis factorial y sus diferentes procedimientos.

c) Evidencia basada en el proceso de respuesta

Este tipo de evidencia refleja la interacción entre la psicología cognitiva y la psicometría, donde el análisis de los procesos cognitivos comprometidos en el proceso de respuesta a los tests adquiere particular importancia.

El análisis empírico y teórico del proceso de respuesta del test puede suministrar evidencia relacionada con la congruencia entre el constructo medido y la naturaleza del rendimiento o respuesta emitida por los examinados (APA, 1999). Por ejemplo, en un test de razonamiento numérico es importante determinar si los examinados están realmente razonando para emitir sus respuestas, en lugar de seguir un algoritmo estándar. En otro

caso, se debería asegurar que los puntajes de una escala que mida extroversión-introversión, por ejemplo, no estén fuertemente influidos por la tendencia hacia la conformidad social.

Esta evidencia generalmente se obtiene por medio de la utilización de entrevistas con los examinados, protocolos de respuesta o cualquier procedimiento que permita el análisis cualitativo de las respuestas individuales a los ítems del test. Los administradores deben examinar las estrategias de respuesta al test utilizadas por los individuos para enriquecer su comprensión del constructo. Esta evidencia puede contribuir a despejar los interrogantes relacionados con las diferencias entre grupos de examinados (mujeres *versus* varones, por ejemplo) en los puntajes de test. Los estudios sobre procesos involucrados en las respuestas de los examinados de diferentes subgrupos (étnicos, por ejemplo) deberían ayudar a esclarecer en qué medida las capacidades irrelevantes o accesorias para el constructo pueden influir diferencialmente en sus rendimientos. Pese a su importancia potencial, en la práctica, esta nueva evidencia de validez (recién contemplada en la última versión de los estándares de la APA) ha sido poco investigada.

Fuentes externas de evidencia

El análisis de las relaciones de las puntuaciones del test con variables externas al mismo test es otra fuente importante de evidencia. Las variables externas pueden ser las medidas de algún criterio que el test intenta predecir, así como las puntuaciones de otros tests que miden constructos semejantes o diferentes.

Las variables categóricas externas (sexo, por ejemplo), incluyendo la pertenencia a grupos diferentes, también son importantes de considerar cuando la teoría sugiere que las diferencias en las puntuaciones de grupos contrastados (esquizofrénicos y no esquizofrénicos, por ejemplo) deberían reflejar diferencias en el constructo medido por el test (APA, 1999). Así, por ejemplo, si la teoría subyacente a un inventario de intereses vocacionales postula que los hombres (como grupo) se interesan por áreas de conocimiento diferentes que aquellas que resultan atractivas a

las mujeres, las puntuaciones de ese test deberían apoyar esa hipótesis demostrando, por ejemplo, que los hombres obtienen puntuaciones medias significativamente más elevadas que las mujeres en la dimensión "Realista" de los intereses y que las mujeres alcanzan puntuaciones significativamente más elevadas que los hombres en la dimensión "Social" de la tipología RIASEC (Holland, 1997).

No obstante, como señaló Hogan (2004), la dificultad principal del método de grupos de contraste es que con una muestra grande es sencillo obtener diferencias estadísticamente significativas entre los grupos comparados. El tamaño de las diferencias debe ser lo suficientemente grande para que el test diferencie entre los grupos en un grado que sea útil en la práctica. En ese sentido, la significación estadística es una condición necesaria pero no suficiente y, por lo tanto, deberían utilizarse índices como *d* (diferencia media tipificada) para evaluar el tamaño del efecto (véase el apartado de generalización de la validez).

Revisemos ahora otras categorías relevantes para la fuente de evidencia de validez.

d) Evidencia convergente-discriminante

Al construir un test (una escala para medir inestabilidad emocional, por ejemplo) se deben comparar los puntajes obtenidos, tanto con otros tests elaborados para medir el mismo atributo (la escala Neuroticismo de un inventario de personalidad, por ejemplo) como con los diseñados para medir otros atributos (la escala Amabilidad del mismo test, por ejemplo). La lógica de ambos procedimientos complementarios es evidenciar que el test en cuestión mide realmente el constructo que se propone medir, al correlacionarse con otros tests reconocidos que miden el mismo constructo, y no correlacionarse con tests que miden constructos diferentes.

Por consiguiente, el test novedoso debe presentar correlaciones significativamente más altas con los tests que miden el mismo rasgo que con aquellos que miden constructos diferentes. La evidencia de convergencia o convergente está dada por correlaciones relativamente altas entre aquellos instrumentos

de medición diseñados para evaluar un rasgo común. Por el contrario, la evidencia de discriminación o discriminante se obtiene cuando se encuentran correlaciones no significativas, muy débiles o negativas entre instrumentos que miden rasgos diferentes. Es claro que "rasgos diferentes" refiere a dimensiones distintas de un mismo constructo (aptitud verbal *versus* matemática o estabilidad emocional *versus* neuroticismo) o constructos relacionados. Por ejemplo, no tendría sentido correlacionar tests de amabilidad y aptitud espacial para obtener evidencia discriminante porque es obvio que esos constructos muy raramente demostrarán correlacionarse.

Una correlación al menos moderada (y significativa) entre los resultados obtenidos por un mismo grupo de examinados en dos subtests de aptitudes cognitivas que midan razonamiento numérico es un ejemplo de validez convergente. En cualquiera de estos subtests la evidencia discriminante se obtendría si las correlaciones, con las puntuaciones de un subtest de comprensión lectora, por ejemplo, fueran negativas, muy débiles o no significativas. La tabla siguiente presenta las correlaciones obtenidas en un estudio de validez de convergencia y discriminación de una escala del NEO-PI-R (Costa y Mc Crae, 1999).

Tabla 4.2. Correlaciones entre la escala Extraversión del NEO-PI-R y otras escalas de tests relacionados

Inventario	Escala	Correlación	Evidencia
SDS	Social	0,67**	Convergente
MMPI-2	Introversión	-42**	Discriminante

** p< 0,01

Como puede apreciarse, la escala Extraversión del NEO-PI-R (Costa y McCrae, 1999) recoge evidencia convergente en relación con escalas que miden rasgos similares, como la escala Social del Self-Directed Search (SDS) (Holland, 1994), y evidencia discriminante respecto a escalas que miden rasgos diferentes,

tal como la escala Introversión del Inventario de Personalidad de Minnesota (MMPI-2) (Butcher y colaboradores, 1989).

Una estimación sistemática de esta evidencia de validez puede realizarse mediante la matriz multirrasgo-multimétodo (Campbell y Fiske, 1954), con la que se obtiene una matriz de las correlaciones entre dos o más instrumentos construidos para medir constructos semejantes. La finalidad esencial de este último procedimiento es demostrar que las correlaciones entre los tests que miden un mismo rasgo (escalas de ansiedad de dos tests diferentes, por ejemplo) son más elevadas que las correlaciones entre constructos diferentes medidos por un mismo test (entre ansiedad y depresión medidas por el MMPI, por ejemplo), y más elevadas aún que las correlaciones entre constructos diferentes medidas por tests diferentes (entre la ansiedad medida por un inventario y la depresión medida por otro). Este método es muy recomendable y está recibiendo atención creciente en la literatura (Hogan, 2004). En la actualidad también se utiliza el análisis factorial confirmatorio (véase el capítulo de construcción de tests) para recoger evidencia convergente-discriminante entre tests.

e) Evidencia de las relaciones entre las puntuaciones del test y criterios externos

Este tipo de evidencia es especialmente importante en los contextos aplicados de la psicología en los que se busca predecir de manera precisa un determinado comportamiento o desempeño a partir de las puntuaciones de un test. En función del número y tipo de variables predictoras y dependientes (criterios), pueden utilizarse diferentes análisis estadísticos dentro del modelo lineal general. En este texto haremos referencia a los procedimientos más utilizados cuando se trata de una variable predictora y un criterio (correlación bivariada) y cuando se utilizan diversas variables para la predicción de un criterio (correlación múltiple). En la actualidad, los diseños de investigación para verificar la utilidad predictiva de un test incluyen casi siempre métodos multivariados, tales como el análisis de regresión múltiple o el análisis de senderos (*path analysis*).

Correlación bivariada con un criterio

En este contexto, la validez de las puntuaciones de un test significa la efectividad con que se puede predecir el desempeño de una muestra en una situación real (laboral o académica, por ejemplo) o criterio diferente del test en sí mismo. El estadístico que se emplea habitualmente para corroborar esa utilidad predictiva es la correlación entre la puntuación obtenida en la prueba y la puntuación en algún criterio que representa los desempeños externos al test.

Los coeficientes utilizados dependen del tipo de variables medidas (dicotómicas, continuas) y de las escalas de medición empleadas (ordinal o intervalar, por ejemplo), siendo algunos de los más usuales el coeficiente producto-momento de Pearson (r) (para dos variables continuas) y el coeficiente punto-biserial (una puntuación dicotómica y otra continua), cuando la escala de medición es intervalar, o el coeficiente de rangos de Spearman, adecuado para el caso de una o dos variables medidas en una escala ordinal (Thorndike, 1989).

Se dispone también de coeficientes especiales que se utilizan cuando las relaciones entre las variables no son lineales sino curvilíneas, tales como el coeficiente eta (Martínez Arias, 1995). Las relaciones entre ansiedad y desempeño, por ejemplo, suelen ser no lineales puesto que un nivel de ansiedad moderado puede favorecer el rendimiento y, en cambio, un nivel elevado interferir en el mismo. En este último caso, si se utilizara un coeficiente que supone relaciones lineales, tal como r de Pearson, se podría subestimar la correlación entre las variables.

En el contexto de la validez, el criterio es una medida directa e independiente de lo que el test intenta predecir o inferir (Martínez Arias, 1995). Para un test de aptitud mecánica, un criterio posible sería el rendimiento posterior en una tarea de reparación de automóviles, así como para una escala que mide psicoticismo, el diagnóstico clínico disponible sería un criterio adecuado.

El diseño del experimento básico que se emplea para verificar si las puntuaciones de un test se relacionan con un criterio es, según Murat (1985), el siguiente:

a. De una población determinada se extrae una muestra grande.
b. Los individuos de la muestra son evaluados tanto con el test como con el criterio que interesa predecir.
c. Con los datos obtenidos del paso *b* se estima un coeficiente de correlación. Si éste es significativamente distinto de cero podrá aseverarse que el test es un predictor del criterio para cualquier individuo que pertenezca a esa población. Si, por el contrario, la correlación no es significativamente distinta de cero es legítimo inferir que las puntuaciones del test carecen de utilidad en relación con ese criterio. Debe recordarse que el nivel de significación (alfa) generalmente se establece en 0,05. Esto implica que existe un 95% de probabilidad de que la correlación observada obedezca a un efecto real y un 5% de probabilidad de que el resultado obtenido en la muestra de investigación sea producto del azar. Expresado de otro modo, la probabilidad de rechazar la hipótesis nula (ausencia de correlación entre las dos variables) cuando ésta es verdadera es solamente del 5%. Este último criterio (significación estadística) debe interpretarse en el contexto de la potencia estadística de una investigación y el tamaño del efecto que se comentará más adelante. En muchos textos de estadística se presenta una tabla de valores críticos que permite interpretar la significación estadística del coeficiente de correlación. No obstante, todos los programas estadísticos usuales (SPSS, por ejemplo) suministran en la actualidad los coeficientes de correlación con el nivel de significación correspondiente.

Para la mayoría de los problemas de predicción es razonable esperar sólo correlaciones moderadas entre un criterio y cualquier test predictor o combinación de tests. Las personas son demasiado complejas como para que sea posible, a partir de resultados de tests, una predicción exacta de su rendimiento e igualmente complejas son las situaciones en las que se obtienen los datos del criterio (Nunnally, 1991).

La tabla siguiente muestra las correlaciones obtenidas entre las puntuaciones en el SAT (un test de aptitudes académicas) y

el promedio de calificaciones anual en una muestra de estudiantes universitarios de los Estados Unidos (Johnson, 1994).

Tabla 4.3. Correlaciones entre el SAT y el rendimiento académico en una muestra de estudiantes universitarios (N = 508)

SAT	Correlación con el promedio de calificaciones anual
Verbal	0,44 **
Cuantitativo	0,37 **

** p< 0,01

Como puede observarse, ambas correlaciones son moderadas (entre 0,30 y 0,49), estadísticamente significativas (nivel de significación 0,01) y demuestran la utilidad predictiva de los subtests del SAT respecto al criterio de rendimiento académico.

Un requisito para los diseños de validez de criterio es que la evaluación de la muestra en la variable a predecir (criterio) debe ser independiente de la evaluación (por ejemplo, dos psicólogos diferentes) en la variable predictora (las puntuaciones del test). La evaluación en el criterio puede ser simultánea (evidencia concurrente) o posterior (evidencia predictiva) a la administración del test (Murat, 1985). Un estudio predictivo obtiene información sobre la precisión con la cual datos procedentes del test pueden ser empleados para estimar puntajes de criterio que serán obtenidos en el futuro y es especialmente pertinente para tests empleados en contextos educativos y ocupacionales. Un estudio concurrente, en cambio, aporta información del criterio simultáneamente a los datos de la prueba y es recomendable para tests elaborados con finalidades de diagnóstico clínico (APA, 1999).

Uno de los problemas más arduos que enfrenta este tipo de evidencia de validez es la selección de los indicadores operacionales del criterio. Legítimamente, podemos interrogarnos acerca de qué constituye el éxito para un estudiante, un mecánico o un vendedor de seguros, por ejemplo. Es bastante difícil formular cualquier definición precisa de éxito en un empleo y mucho

más desarrollar una medida que lo represente adecuadamente. Como plantea Thorndike (1989), la historia completa de la efectividad de una persona en un dominio tiene lugar en el curso de toda una vida, no se verifica en un momento específico, y aunque se pueda anticipar un cierto grado de consistencia en el tiempo, claramente una sección transversal es una imagen parcial del éxito de un individuo en un dominio particular.

Medir el rendimiento académico, por ejemplo, parece sencillo puesto que las personas generalmente son calificadas en forma cuantitativa y esa información (notas, promedios) está casi siempre disponible para el investigador. Sin embargo, debe reconocerse que las calificaciones académicas son solamente un indicador parcial (y algunas veces poco confiable) del logro de una persona en cualquier nivel educativo. La estructuración y el desarrollo de criterios de calificación adecuados con los que evaluar la validez de pronóstico de las puntuaciones de tests es uno de los aspectos que requieren mayor reflexión por parte del investigador. En las obras citadas de Thorndike (1989) y Martínez Arias (1995) se realiza un análisis detallado de los indicadores más comunes según el criterio que se desee predecir y de sus principales limitaciones.

Por otra parte, existen factores que pueden afectar las relaciones entre los puntajes de un test y las medidas del criterio. Tres de los más importantes son:

- *Diferencias de grupo*: se refieren a variables que pueden influir (moderar) en las correlaciones entre un test y un criterio, tales como sexo, edad o nivel socioeconómico. En general, se recomienda emplear una muestra lo más heterogénea posible. Debe tenerse en cuenta que la magnitud de los coeficientes de validez es más reducida en grupos más homogéneos, con un rango más limitado en las puntuaciones de pruebas. Puesto que la magnitud de un coeficiente de correlación es una función de dos variables, la reducción del rango de calificación, ya sea en la variable predictora o en el criterio, tenderá a disminuir el coeficiente de validez.
- *Variabilidad de las respuestas al test y al criterio:* en este caso, los puntajes de un test extenso poseen varianzas mayores y las puntuaciones obtenidas en pruebas breves tienen varianzas más reducidas y, por consiguiente, coeficientes de correlación inferiores con las medidas del criterio. Estas afirmaciones deben entenderse atendiendo a los condicionantes formulados en el capítulo de confiabilidad, en el sentido de que la calidad de los ítems es un factor más relevante que la mera cantidad. Se ha demostrado (Pajares, Hartley y Valiante, 2001) que los tests con varias alternativas de respuesta también incrementan la variabilidad de respuesta. Una escala *likert* puede producir igual o mayor variabilidad de respuesta que una prueba extensa con formato dicotómico, tal como correcto-incorrecto.
- *Confiabilidad del test y del criterio*, puesto que en la medida en que el test predictor y el criterio aumenten su confiabilidad se incrementará el coeficiente de validez test-criterio. Por el contrario, una merma en la confiabilidad del test o el criterio van a disminuir el coeficiente de validez. De hecho, el tamaño del coeficiente de validez no puede superar la raíz cuadrada del producto entre los dos coeficientes de confiabilidad (del predictor y del criterio).

En general, la magnitud de los coeficientes de validez no es tan elevada como sería esperable, debido a la multideterminación del comportamiento humano y a las limitaciones de los instrumentos de medición empleados en psicología, analizadas en la primera parte de este libro. Valores significativamente distintos de cero y cercanos a 0,30 ya son atendibles.

En el lenguaje estadístico contemporáneo, los coeficientes de correlación son considerados como una medida del tamaño del efecto, que informa sobre la importancia de los resultados de una investigación. Las pruebas de significación estadística nos previenen del error tipo 1 (o alfa), vale decir, la probabilidad de rechazar la hipótesis nula (que expresa que no hay diferencias o no hay relación entre dos variables) cuando ésta es verdadera. De este modo, establecer un nivel de significación de 0,05 nos permite afirmar que sólo existe una probabilidad del 5% de que el resultado observado en la muestra de investigación sea producto del azar, pero nada nos dice sobre la magnitud de la relación o el efecto de una intervención. Atendiendo sólo al cri-

terio de significación estadística es probable que un coeficiente de correlación débil sea significativo y uno fuerte aparezca como no significativo, puesto que el nivel de significación de un estadístico depende del tamaño de la muestra. Por consiguiente, para tener una perspectiva más integral de los resultados de una investigación, la información sobre la significación estadística debe complementarse con la del tamaño del efecto.

Los coeficientes de correlación entre 0,10 y 0,29 sugieren una magnitud pequeña de la relación; de 30 a 0,49 moderada, y de 0,50 o superiores, un tamaño del efecto grande (Aron y Aron, 2001). Otro aspecto importante relacionado con el nivel de significación y el tamaño del efecto es la potencia estadística de una investigación. La potencia nos permite prevenir el riesgo de aceptar la hipótesis nula cuando es falsa (error tipo II o beta), y depende del tamaño del efecto pero también del tamaño muestral. Para que la potencia estadística de una investigación donde se informa un coeficiente de correlación moderado entre dos variables sea del 80% (el nivel de potencia que se considera aceptable) es necesaria una muestra grande (100 individuos aproximadamente). Por su parte, un tamaño del efecto grande sólo requiere una muestra pequeña (N = 30) para que la potencia estadística de la investigación sea razonable. Por el contrario, un tamaño del efecto débil (r = 0,2, por ejemplo) necesita aproximadamente de 800 participantes para que la potencia estadística sea del 80%. Los tests de inteligencia (uno de los mejores predictores del arsenal psicométrico) demuestran correlaciones promedio de 0,45, moderadas-altas, con indicadores de rendimiento académico tales como el promedio anual de calificaciones, en estudios meta-analíticos y con muestras muy extensas (Jensen, 1998).

También es importante considerar el coeficiente de determinación (coeficiente de correlación al cuadrado), que nos indica la proporción de la variabilidad del criterio que podemos predecir legítimamente a partir de la información del test predictor. De esta manera, un coeficiente de 0,50 indica que las puntuaciones del test predictor nos permite explicar un 25% de la varianza del criterio, mientras que el 75% restante debería atribuirse a la influencia de variables no contempladas por el test.

Si se correlacionan dos variables, un predictor y un criterio,

el grado de asociación entre ellas puede ser influido o moderado por otras variables. La correlación bivariada entre motivación y rendimiento académico en una muestra podría ser de 0,35, pero descender a 0,20 cuando se controla la influencia de la inteligencia (se mantiene artificialmente constante) sobre ambas variables o alguna de ellas. En otro ejemplo, un investigador podría descubrir una asociación fuerte entre estrés marital y tiempo de matrimonio (a más tiempo de convivencia marital, mayor estrés) pero sospecha que esa relación puede ser influida por el número de hijos de la pareja. Es decir, a mayor cantidad de hijos mayor estrés y, a su vez, el tiempo de matrimonio se relaciona generalmente con el número de hijos de la pareja. Por consiguiente, el estrés podría ser, en parte o totalmente, ocasionado por el número de hijos y no tanto por el tiempo de convivencia. Para verificar esta última hipótesis sería necesario controlar (eliminar, mantener constante) el efecto de la variable número de hijos para verificar cuál es la contribución explicativa independiente de la variable tiempo de matrimonio a la variabilidad del estrés percibido en la relación. Del mismo modo, si se quisiera verificar la contribución independiente de la variable cantidad de hijos deberíamos controlar el efecto del tiempo de matrimonio, manteniendo sus valores constantes para toda la muestra (Aron y Aron, 2001).

Este control de las variables intervinientes (siempre presentes en el comportamiento humano) es factible mediante los métodos multivariados de investigación. Estos métodos (así como el meta-análisis que se revisará más adelante) permiten obtener una estimación más poderosa y estable de las magnitudes reales de las correlaciones entre variables predictoras y criterios.

Correlación múltiple con un criterio

Actualmente, la mayoría de las investigaciones realizadas en psicología y ciencias sociales utilizan un enfoque multivariado puesto que permite esclarecer las interrelaciones entre un conjunto de predictores y uno o más criterios, y no meramente verificar el grado de asociación entre un predictor y un criterio como en la correlación bivariada (Thompson y Borrello, 1985).

Los métodos multivariados son numerosos (análisis factorial, regresión múltiple, análisis de senderos, análisis discriminante, entre otros) y se emplean con diferentes finalidades, tales como predecir la pertenencia a un grupo, explicar la variabilidad de una variable dependiente o verificar la estructura de un constructo medido.

En el contexto de la validez de los tests, mediante métodos multivariados como el análisis de regresión múltiple podemos estimar en cuánto se incrementa la precisión de la predicción de un criterio cuando un test se incluye en una batería de tests en comparación con las ocasiones en que no se lo incluye (Aiken, 2003). Por ese motivo, la evidencia de validez obtenida mediante un procedimiento multivariado (como el análisis de regresión múltiple o el análisis de senderos) también se denomina incremental.

El análisis de regresión múltiple permite ponderar la contribución independiente realizada por cada variable predictora para la explicación de un criterio determinado, así como estimar la contribución conjunta de un conjunto de predictores a la explicación del criterio (Aron y Aron, 2001). La ecuación de regresión múltiple se basa en la correlación de cada test con el criterio, pero también informa sobre las correlaciones entre los tests predictores. Las pruebas que correlacionan más alto con el criterio reciben más peso en la ecuación aunque es igualmente importante considerar la correlación de cada prueba con las restantes. Los tests que correlacionan alto entre sí representan una duplicación innecesaria puesto que explican casi los mismos aspectos del criterio (Anastasi y Urbina, 1998).

Los estadísticos fundamentales del análisis de regresión múltiple son: a) el coeficiente de regresión estandarizado o beta (β), que indica cual es la importancia relativa de cada variable independiente en la predicción de la variable dependiente; b) el coeficiente de correlación múltiple (R) que expresa el grado de asociación entre dos o más variables independientes (predictoras), en conjunto, con una variable dependiente; c) el coeficiente de corrrelación múltiple al cuadrado (R^2), que permite determinar el porcentaje de varianza explicada de la variable dependiente por parte del conjunto de predictores de la ecuación, y d) el cambio en R^2 que indica el porcentaje de varianza de la variable dependiente explicada independientemente por

cada uno de los predictores. Estadísticos adicionales como el análisis de la comunalidad y los coeficientes de correlación semiparcial son también muy útiles para interpretar los resultados del análisis de regresión múltiple, refinando la comprensión de la varianza explicada específica de cada predictor sin estar contaminada por la varianza común (compartida por los predictores). En el apéndice se suministra un ejemplo completo de un análisis de regresión múltiple.

Lo que incrementa el coeficiente de correlación múltiple al añadir una variable predictora a la ecuación de regresión es precisamente la correlación semiparcial de esa variable predictora (inteligencia, por ejemplo), con el criterio (rendimiento académico, por ejemplo) controlando la influencia de las restantes variables predictoras (la motivación, por ejemplo) sobre ese predictor (inteligencia, en este caso) (Muñiz, 2001). Como explicamos más arriba, controlar estadísticamente una variable significa mantener constantes sus valores. En el ejemplo anterior, para estimar la contribución específica de la motivación para el rendimiento académico los valores de la muestra de investigación en la medida de la inteligencia deben ser constantes. El razonamiento implícito es: si todos los sujetos fueran igualmente inteligentes, ¿cuánto contribuye la motivación a explicar el rendimiento académico? Recapitulando, el coeficiente de correlación semiparcial indica el grado de asociación existente entre la variable dependiente y la parte de la variable independiente que no está explicada por el resto de las variables independientes de la ecuación de regresión.

Existen varios métodos de regresión múltiple, pero los tres más empleados son: el análisis de regresión jerárquica o secuencial, la regresión estándar y la regresión *stepwise* (paso por paso) (Tabachnick y Fidell, 2001).

En el análisis de regresión jerárquico, las variables predictoras son ingresadas a la ecuación de predicción en el orden lógico sugerido por la teoría y es el procedimiento más recomendable. Así, por ejemplo, si un investigador quisiera predecir el rendimiento académico utilizando tests de aptitudes y de autoeficacia, las medidas de aptitudes deberían ser ingresadas primero, puesto que son una de las fuentes de la autoeficacia y, posteriormente, las medidas de este último constructo.

El método *stepwise* se aplica en las fases exploratorias de investigación, cuando no existe un modelo teórico explícito previo. En este método, el programa de computación (SPSS, por ejemplo) selecciona primero la variable independiente que presenta la correlación significativa más elevada con la dependiente. El proceso continúa hasta que la incorporación de variables al modelo no implique una mejora significativa en la predicción (Aron y Aron, 2001). Este método posee limitaciones que lo hacen poco recomendable excepto con fines exploratorios o para encontrar la mejor ecuación de predicción de una variable sin importar su significado teórico. Una de las principales dificultades es que los resultados son muy dependientes de las características de la muestra empleada en una investigación determinada. En efecto, el orden de ingreso de las variables que determina el programa puede depender de diferencias triviales en las relaciones entre los predictores en una muestra en particular que no reflejan diferencias reales en la población (Tabachnick y Fidell, 2001).

Cuando no se cuenta con una base teórica sólida que permita inferir el orden de ingreso de las variables de manera secuencial, es preferible utilizar el método de regresión estándar, en el que todas las variables predictoras son ingresadas en un mismo paso y luego se analiza cuál es la contribución específica de cada una a la predicción de la variable dependiente.

El análisis de regresión múltiple, pese a su relativa sencillez (comparado con otros métodos multivariados), posee una serie de supuestos exigentes que el investigador debe respetar para interpretar inequívocamente los resultados. Entre ellos cabe mencionar por su importancia: el tamaño muestral requerido (104 casos más el número de variables independientes incluidas en el modelo como regla general), el análisis de los casos con valores extremos en las variables (casos marginales), la ausencia de multicolinearidad (correlaciones elevadas entre los predictores), la distribución normal de todas las variables, la linealidad de las relaciones entre las variables y la adecuada confiabilidad de las medidas empleadas (Tabachnick y Fidell, 2001).

En el análisis de regresión múltiple, la estimación aproximada del tamaño del efecto puede realizarse elevando al cuadrado

el coeficiente de correlación múltiple. Así, por ejemplo, un R^2 de 0,25 (R = 0,50) indicaría que estamos explicando un 25% de la varianza del criterio (rendimiento académico en matemática, por ejemplo) con las variables predictoras (inteligencia, autoeficacia y rendimiento anterior en matemática, por ejemplo) del modelo de predicción. Este último valor es equivalente a una diferencia media tipificada (d) de 1 y sugiere un tamaño del efecto grande. Con más precisión, en el contexto de la regresión múltiple se utiliza el estadístico f^2 de Cohen (1988) como estimador del tamaño del efecto. Los valores de f^2 entre 0,02-0,14 indican un tamaño del efecto pequeño, entre 0,15-34 moderado y 0,35 o superiores un efecto grande. La fórmula para estimar f^2 en la regresión múltiple estándar es:

$$f^2 = \frac{R^2}{1 - R^2}$$

Donde R^2 es el coeficiente de correlación múltiple al cuadrado.

Los diferentes métodos del análisis de regresión múltiple son potentes para estimar la explicación de una variable dependiente por parte de un conjunto de predictores, pero no tanto para comprender las interrelaciones entre las variables independientes incluidas en el modelo. Estas limitaciones son subsanadas con el empleo del análisis de senderos (*path analysis*), un método especial del modelo de ecuaciones estructurales que permite comprender con más claridad la red de intercorrelaciones entre las variables y determinar no sólo las contribuciones directas a la explicación de una variable, sino también las indirectas.

Cuando la variable a predecir o criterio es nominal, por ejemplo, la pertenencia a una ocupación o carrera, es adecuado emplear otros métodos multivariados, tales como la regresión logística o el análisis discriminante múltiple. No obstante, la interpretación de los resultados es compleja y, al igual que en el análisis de senderos, la descripción de estos métodos no es pertinente para los propósitos de este texto introductorio. La obra de Tabachnick y Fidell (2001) es una de las mejores referencias para el estudio de la metodología multivariante de investigación.

f) Evidencia de las consecuencias de la aplicación de tests

En los últimos años, se ha prestado atención a las consecuencias esperadas e inesperadas de la aplicación de tests, aunque es uno de los conceptos novedosos más controvertidos en relación con la validez (APA, 1999; Oliden, 2003).

Es importante diferenciar la evidencia que es directamente relevante para la validez de aquella que sólo tiene significación para las decisiones en políticas sociales. En ese sentido, existe interés por conocer las causas de las diferencias grupales observadas en los puntajes de tests en ámbitos tales como la selección y promoción laboral o la colocación de niños en clases de educación especial. Aunque la información acerca de las consecuencias de la aplicación de tests puede influir sobre las decisiones a propósito del empleo de tests, tales consecuencias no afectan directamente la validez de las interpretaciones de las pruebas. Los juicios de validez o invalidez de los tests en relación con las consecuencias de su administración dependen de una investigación más minuciosa sobre las fuentes de tales consecuencias (APA, 1999).

Considérese, por ejemplo, el caso de las puntuaciones observadas en miembros de grupos diferentes (étnicos, por ejemplo) como consecuencia del uso de tests en psicología laboral. Si estas diferencias son ocasionadas solamente por una desigual distribución de las destrezas que el test mide, y si estas habilidades son, en realidad, predictores importantes del desempeño laboral, entonces el descubrimiento *per se* de las diferencias intergrupales no implica una falta de validez de este test. Si, por el contrario, el test mide diferentes destrezas no relacionadas con el desempeño (por ejemplo, un nivel sofisticado de lectura para un trabajo que sólo exige capacidades mínimas en este dominio) o si las diferencias se deben a la sensibilidad del test para algunas características de los examinados que no se espera formen parte del constructo medido, entonces la validez de estos resultados debería ser revisada, aunque los puntajes correlacionen positivamente con algún criterio de rendimiento ocupacional. En este sentido, la determinación de fuentes de sesgo en las puntuaciones de tests (que revisamos en el último capítulo) es muy pertinente para este tipo de evidencia de validez.

En síntesis, la evidencia relacionada con las consecuencias de la aplicación de tests sólo es relevante para la validez cuando se relaciona con alguna fuente de invalidez, tales como una pobre representación del constructo o la existencia de componentes no relevantes para el constructo. Al igual que acontece con la evidencia de proceso de respuesta al test, la evidencia de consecuencias es poco investigada en la práctica.

4.3. Utilidad de los tests en contextos de clasificación

Como señala Thorndike (1989), el coeficiente de correlación entre una variable predictora y un criterio externo es un estadístico necesario pero insuficiente cuando se trata de determinar el valor práctico de un test como herramienta para la toma de decisiones en situaciones que demandan la clasificación de personas. Para Cronbach (1998), los tests pueden usarse con cuatro propósitos diferentes:

1. Autoconocimiento
2. Clasificación
3. Evaluación de programas de intervención
4. Investigación

El autoconocimiento facilitado por tests sirve primordialmente al individuo, y sólo marginalmente a las instituciones. Éste sería el caso de los tests empleados en orientación, tal como un inventario de intereses vocacionales cuyos resultados son útiles para la elección de carrera de la persona examinada. Por el contrario, el segundo propósito, clasificación, se relaciona con decisiones que afectan a las personas a las que se les administra el test. La clasificación sirve principalmente a las instituciones y se produce cuando cualquier persona es asignada a una categoría y no a otra. El propósito de clasificación se presenta en situaciones de selección o diagnóstico, tales como exámenes aplicados a aspirantes al ingreso a un empleo o asignación de los individuos a una categoría diagnóstica, como depresión o fobia. Es decir, en situaciones donde el criterio que se intenta predecir es dicotómico, por ejemplo, aprobar-reprobar, enfermo-no enfermo, admitido-

rechazado (Muñiz, 2001). En todas estas situaciones de clasificación establecer la validez de las puntuaciones de un test mediante un coeficiente de correlación es claramente insuficiente.

La validez o utilidad real de las puntuaciones de un test en situaciones de clasificación se demuestra comprobando que existe una diferencia real entre las predicciones efectuadas a partir de esas puntuaciones y las que sería posible efectuar a partir del mero conocimiento del comportamiento promedio de todos los individuos en la variable de interés (Murat, 1985).

La lógica del proceso de validación en este contexto es clasificar a las personas en dos categorías (alumnos exitosos y no exitosos en el primer año de una carrera, por ejemplo) y dicotomizar las puntuaciones en el test predictor (examen de ingreso a una carrera, para este caso) a partir de determinado punto de corte (calificación de 7 puntos, por ejemplo). Las clasificaciones realizadas a partir de los puntajes obtenidos por las personas en el test predictor (el examen de ingreso) deberían ser coincidentes con las realizadas a partir del desempeño en el criterio (Muñiz, 2001). Esta coincidencia entre las predicciones efectuadas a partir del test y el desempeño real de las personas en el criterio puede verificarse mediante varias estrategias, algunas de las cuales se expondrán a continuación.

Un caso típico sería, por ejemplo, un test referido a criterio utilizado para seleccionar empleados que desean ingresar a una ocupación. Para determinar si es válido para ese propósito de clasificación se administra el test al grupo de aspirantes al empleo y se permite ingresar a todo el grupo, independientemente de la puntuación obtenida en el test. Al finalizar un período de trabajo se evalúa el rendimiento de todo el grupo de ingresantes. En el test predictor pueden fijarse arbitrariamente varios "puntos de corte" o puntos críticos con respecto a los cuales podríamos eventualmente efectuar nuestros pronósticos de éxitos o fracasos (una calificación de 7 en el test, por ejemplo).

Como resultado de la aplicación de este procedimiento se obtienen cuatro resultados posibles:

1. "A": Individuos pronosticados como "éxitos" y que resultaron serlo (aciertos positivos). Aceptados por el test y con rendimiento adecuado en el criterio.

2. "C": Individuos pronosticados como "fracasos" y que, efectivamente resultaron serlo (aciertos negativos). Rechazados en el test y con bajo rendimiento en el criterio.
3. "B": Individuos pronosticados como "éxitos" y que, por el contrario fracasaron (falsos positivos). Aceptados por el test y con bajo rendimiento en el criterio.
4. "D": Individuos pronosticados como "fracasos" y que luego resultaron ser exitosos (falsos negativos). Rechazados por el test y con alto rendimiento en el criterio.

Si colocásemos valores numéricos a una tabla de contingencia de este tipo, tendríamos, por ejemplo:

ACIERTOS POSITIVOS	**FALSOS POSITIVOS**
5	2
ACIERTOS NEGATIVOS	**FALSOS NEGATIVOS**
4	1

- *Falsos positivos*: (2) son dos las personas de las cuales el test predice que tendrán buen rendimiento pero no lo obtienen en el criterio.
- *Falsos negativos*: (1) según las puntuaciones del test la persona tendría mal rendimiento, pero en el trabajo se desempeña bien.
- *Aciertos positivos*: (5) cinco son las personas para las cuales la predicción del test resultó correcta, vale decir, tuvieron buen rendimiento en el test y en el criterio.
- *Aciertos negativos*: (4) serían cuatro las personas para las cuales la predicción del test también resultó acertada, es decir, tuvieron mal rendimiento en el test y en el criterio.

El índice más sencillo sería calcular la proporción de clasificaciones correctas realizadas a partir de las puntuaciones del test o poder predictivo total del test, que con estos datos estaría dado por la sumatoria de los aciertos (positivos + negativos) sobre el total de observaciones:

En el ejemplo anterior:

PPT (poder predictivo total) = 5+4/12 = 0,75

Esto implicaría que el test clasifica correctamente a los individuos del ejemplo en un 75% de los casos. El coeficiente kappa (examinado en el capítulo 3) es un índice más conservador, atenuando el valor del PPT al restar de los aciertos aquéllos debidos al azar.

Un ejemplo real de validación de un test en contextos de clasificación puede observarse en una investigación realizada con la Escala de Evaluación de la Demencia de Mattis (Fernández y Scheffel, 2003). En este caso se fijó un punto de corte de 123 puntos para la escala (un puntaje menor a éste indicaría la presencia de demencia) luego de un examen de la literatura sobre este tema. Aplicado el instrumento a 60 individuos, se verificó que este punto de corte clasificaba a 5 sujetos como dementes cuando en realidad no lo eran (falsos positivos) y a 3 como no dementes cuando en realidad lo eran (falsos negativos). De este modo, un 13,33% de la muestra era incorrectamente clasificado, tal como se ve en la tabla que sigue. Comparados estos resultados con otros puntos de corte sugeridos en la literatura, se observó, por ejemplo, que los puntajes de corte 127 y 137 aumentaban considerablemente el porcentaje de sujetos erróneamente clasificados.

Tabla 4.4. Evaluación de Demencia según Escala de Mattis. Aciertos y errores en la predicción

Según el test	**Situación real (diagnóstico psiquiátrico)**	
	Dementes	No dementes
Dementes	44 (aciertos positivos)	5 (falsos positivos)
No Dementes	3 (falsos negativos)	8 (aciertos negativos)

Con estos datos puede estimarse la sensibilidad, especificidad y poder predictivo total del instrumento en ese punto de corte.

La sensibilidad (SEN) es la capacidad del instrumento para identificar correctamente al grupo criterio, en este caso las personas con demencia (Hogan, 2004), donde SEN = AP (Aciertos positivos) / AP + FN (Falsos negativos),

SEN = 44 / 44 + 3= 44 / 47 = 0,93

La especificidad (SPE) es la capacidad del test para identificar correctamente al grupo contraste, en este caso los individuos sin demencia, donde SPE = AN (Aciertos negativos) / AN + FP (Falsos positivos),

SPE = 8 / 8 + 5 = 8 / 13 = 0,61

El poder predictivo total del instrumento (PPT), o capacidad global del instrumento para clasificar correctamente los casos, estaría dada en este ejemplo por:

PPT = 44 + 8 / (44 + 5 + 3 + 8) = 52 / 60 = 0,86

Ahora bien, en cualquier situación de este tipo se coloca un punto de corte arbitrario en las puntuaciones del test. Cabe interrogarse acerca de las razones para establecer ese valor como punto de corte y no otro. En efecto, ¿qué errores hubiéramos cometido si fijamos otro punto de corte en las puntuaciones de la variable predictora? Una posible solución a este problema estaría dada por el análisis de los índices mencionados anteriormente (sensibilidad, especificidad, poder predictivo total) en diferentes puntos de corte para determinar cuál produce mejores resultados de clasificación, tal como se comentó en el ejemplo previo.

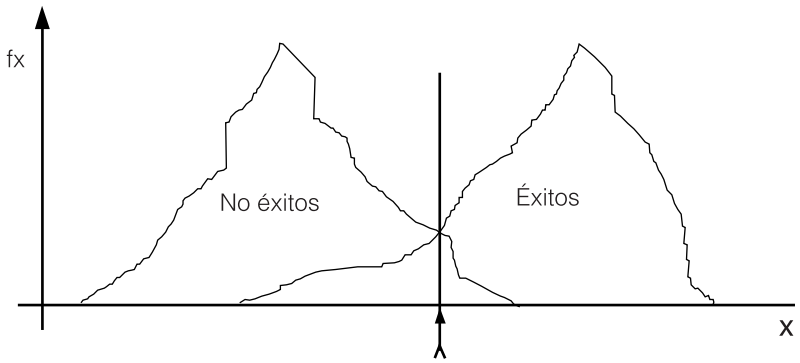
Naturalmente, la determinación de un punto de corte va a depender también de la importancia acordada a los dos tipos de errores: falso positivo y negativo. Todo sistema de previsión del comportamiento humano implica siempre la presencia de un "error", ya sea de tipo "falso positivo" o "falso negativo". Despla-

zando el punto de corte en la variable predictora es posible hacer igual a 0 uno de los dos errores, pero en ningún caso se lo podrá hacer con ambos errores al mismo tiempo.

Según Murat (1985), una solución alternativa, siempre y cuando pueda ser justificada, es la que considera el costo de los dos errores como si fuera igual (da lo mismo equivocarse en un sentido que en el otro). En este caso, si se puede sostener este supuesto (igual costo de ambos errores), la determinación del "punto de corte óptimo" podría obtenerse de este modo:

a) Se aplica el test (para seleccionar pilotos, por ejemplo) a una muestra representativa de la población en estudio.
b) Se observa y se registra el resultado de todos los sujetos así examinados en la variable criterio (simulador de vuelo, por ejemplo).
c) se divide el grupo en dos partes, según si tuvieron "éxito" o si "fracasaron" en la variable criterio.
d) Se construyen dos polígonos de frecuencia (uno con el grupo que tuvo éxito y otro con el que fracasó en la variable criterio) para los valores que cada uno de estos dos grupos registra en la variable predictora.

Figura 4.1. Punto de corte que minimiza los errores en la predicción



El punto de corte que corresponda a la intersección de los dos polígonos, tal como se ve en la figura anterior, hace mínima la suma total de los dos errores, puesto que se asumió que el costo unitario era igual para ambos.

En la mayoría de los casos, sin embargo, el punto de corte deberá establecerse después de una cuidadosa consideración de los costos y de las consecuencias humanas y sociales de cada uno de los dos errores. En ciertas situaciones deberá establecerse un punto de corte lo suficientemente elevado para disminuir al mínimo la posibilidad de fracasos; éste sería el caso de un trabajo en que un empleado erróneamente evaluado (por ejemplo, un piloto de líneas comerciales) pueda causar daños o pérdidas graves. En esta última circunstancia, sería preferible reducir al mínimo los falsos positivos (los que superan el test pero fracasan en el criterio) y, por consiguiente, aumentar los falsos negativos (los que fracasan en el test pero están calificados para volar). En otras circunstancias probablemente se pretenda reducir al mínimo la posibilidad de cometer errores falsos negativos (personas con probabilidades de suicidarse pero con bajas puntuaciones en un test de tendencias suicidas, por ejemplo) (Hogan, 2004).

Un concepto relacionado con este tema, de mucha importancia en psicología clínica y epidemiología, es el de tasas básicas. Una tasa básica se refiere a la frecuencia (expresada en porcentaje) de una condición patológica en una población determinada. Por ejemplo, si el 5% de la población admitida en un hospital padece lesión cerebral, la tasa base de lesión cerebral en esa población es del 5% (Anastasi y Urbina, 1998). Es evidente que la probabilidad de que un miembro cualquiera de esa comunidad padezca lesión cerebral es igual a 0,05.

Desde la perspectiva de la teoría de la decisión (Cronbach y Glaser, 1972), las tasas básicas representan las proporciones de "aciertos" que sería posible lograr prescindiendo de la información proporcionada por el test predictor. Para poder conocer la utilidad predictiva real de un test en este tipo de situaciones sería necesario comprobar que éste mejoraría la capacidad predictiva o de diagnóstico (porcentaje de aciertos) de las tasas básicas (empleadas exclusivamente como predictoras). No obstante, cuando las tasas básicas son muy extremas (muy elevadas o

muy bajas), resulta dificultoso demostrar que las puntuaciones de un test evidencian utilidad predictiva para identificar a los individuos de esa población de interés.

Algunos métodos de investigación multivariantes, tales como el análisis de discriminación y el análisis de regresión logística, también son útiles en situaciones de clasificación. En efecto, estos métodos permiten apreciar la validez de un test para clasificar correctamente a un grupo de personas en un criterio en comparación con la proporción de clasificaciones correctas que hubiéramos obtenido por mero azar. En la obra ya mencionada de Tabachnick y Fidell (2001) puede apreciarse un tratamiento exhaustivo de estos métodos, así como ejemplos de aplicación en contextos de clasificación.

4.4. Generalización de la validez: el meta-análisis

Como expresamos reiteradamente, la validación de un test está vinculada a la muestra particular que se ha empleado en ese proceso. Esta importante limitación intenta ser atenuada por los procedimientos de generalización de la validez. El más empleado en la actualidad es el *meta-análisis*, método que permite integrar y combinar los resultados de diversos estudios empíricos mediante técnicas específicas. En lo que se refiere a la validez, la utilización de este método surgió como respuesta a los coeficientes débiles de correlación observados en muchas investigaciones en relación con la evidencia de las relaciones test-criterio (Martínez Arias, 1995).

Los procedimientos de generalización de la validez comenzaron a aplicarse durante la década de los ochenta, y ya en 1985 las Normas Técnicas de la APA describen al meta-análisis como un procedimiento óptimo para ese propósito.

El meta-análisis es útil para este propósito de generalización al agregar estudios bien diseñados pero obstaculizados por una muestra pequeña y al promediar efectos falsos que operan en ambas direcciones (Nunnally y Bernstein, 1995). Es considerado en la actualidad como una metodología potente de investigación que permite integrar los resultados de diversos estudios empíricos y sus descubrimientos particulares.

Tal como afirman algunos autores (Abelson, 1998; Gómez Benito, 1987; Glass, McGaw y Smith, 1981), hasta hace dos décadas el problema de los resultados contradictorios en investigaciones sobre una misma problemática era resuelto por medio de revisiones narrativas o de resúmenes verbales que presentaban aquellos estudios favorables a una hipótesis particular, las investigaciones que la refutaban y aquéllas que presentaban resultados mixtos. Esta metodología fue criticada por ser demasiado subjetiva e informal y se pensó en el meta-análisis como una alternativa potencialmente superadora. En la actualidad se prefiere este último método frente a las revisiones narrativas tradicionales que se realizan en el marco teórico o antecedentes de una publicación científica (Gómez Benito, 1987).

El meta-análisis transforma los resultados estadísticos de investigaciones empíricas independientes a una métrica común, provee una estimación simple de la fortaleza de la relación entre determinadas variables y permite comprobar estadísticamente si una serie de estudios, conjuntamente considerados, apoyan o refutan las hipótesis de investigación (Multon, Brown y Lent, 1991).

Este método comienza con la reunión, clasificación y codificación de las investigaciones existentes sobre un tema. Dicho proceso implica la consideración, clasificación y codificación de las características sustantivas y metodológicas de las investigaciones particulares (tales como tipo y duración de una intervención o tratamiento experimental, tipo de muestra e instrumentos empleados). El propósito de la codificación de los estudios particulares es verificar si los resultados difieren en función de las características de los mismos.

Debido a que los resultados de las investigaciones podrían ser difíciles de comparar directamente, se los debe transformar en una medida común. Las dos medidas más utilizadas para cuantificar e integrar los resultados de las investigaciones independientes son los niveles de significación y las medidas de tamaño del efecto. El nivel de significación informa si los resultados obtenidos han ocurrido probablemente por azar, mientras que el tamaño del efecto indica la intensidad de la relación o del efecto de interés (Gómez Benito, 1987).

Existen numerosas medidas del tamaño del efecto, pero en la práctica las más utilizadas son la diferencia media tipificada

(*d*) y los coeficientes de correlación (*r*). Una vez transformadas las unidades de análisis de un meta-análisis a una escala común, se aplican técnicas estadísticas específicas que permiten resumir los resultados particulares de las diferentes investigaciones en un índice global.

Los procedimientos utilizados para generalizar las evidencias de validez son básicamente métodos de meta-análisis aplicados a estudios correlacionales, ya que la correlación es el estadístico utilizado en los diseños de investigación que pretenden proporcionar evidencias de validez acerca de las relaciones test-criterio. También puede emplearse el meta-análisis para evaluar otros aspectos de la validez. Así, por ejemplo, cuando se analiza la evidencia convergente de un test con otro semejante, el meta-análisis podría ser un medio idóneo para recoger y sintetizar los estudios correlacionales particulares en relación con ese tipo de evidencia.

No obstante, el meta-análisis se utiliza más frecuentemente para generalizar la evidencia predictiva de un test en relación con un criterio. Por ejemplo, se han realizado meta-análisis con investigaciones empíricas de las relaciones entre las puntuaciones del test de inteligencia para adultos WAIS (Wechsler, 1999) y el rendimiento académico (Hogan, 2004).

El meta-análisis también puede emplearse para obtener evidencia de las relaciones hipotetizadas entre constructos y criterios evaluados por diferentes instrumentos. Este uso más teórico del método se ejemplifica en la investigación clásica de Multon, Brown y Lent (1991). Estos autores condujeron dos investigaciones meta-analíticas con el objetivo de comprobar si diferentes medidas de autoeficacia se relacionaban con el rendimiento académico y con la persistencia académica. Para la realización de ambos meta-análisis se utilizó una muestra total de 39 investigaciones empíricas. Como medida del tamaño del efecto se utilizó el coeficiente de correlación entre autoeficacia y rendimiento en el primer estudio y entre autoeficacia y persistencia en el segundo. Los tamaños del efecto globales observados (0,38 entre autoeficacia y rendimiento, y 0,34 entre autoeficacia y persistencia) corroboraron la existencia de una correlación positiva y moderada entre la autoeficacia y las variables criterio rendimiento y persistencia, respectivamente.

Aunque el meta-análisis es un método de gran utilidad y muy empleado (aunque no es el caso de la Argentina, donde lamentablemente se aplica poco), no se ha visto libre de críticas. No obstante, el avance en los desarrollos estadísticos así como el perfeccionamiento del método han permitido la superación de la mayoría de estos obstáculos.

Uno de las críticas más frecuentes atañe a los sesgos de publicación, es decir, el sesgo de selección editorial a favor de estudios con resultados significativos que favorezcan las hipótesis de investigación. El problema de los sesgos de publicación ha sido enfrentado de diferentes maneras. Por una parte, la literatura especializada recomienda incluir en los meta-análisis estudios sin publicar, tales como disertaciones. En la actualidad este acceso a material inédito se ve facilitado por muchas bases de datos internacionales *on line*. De esta manera, el investigador puede comparar los resultados de los estudios publicados *versus* los no publicados y estimar la existencia de sesgo de publicación. Otra forma de enfrentar este problema es estimando el número de estudios adicionales con resultados no significativos (es decir, que apoyan la hipótesis nula) que serían necesarios para revertir la conclusión de que existe una relación significativa surgida de un meta-análisis (Rosenthal y Di Matteo, 2001).

Otra crítica que habitualmente se hace al meta-análisis es que sus resultados muchas veces no son interpretables debido a que se incluyen estudios poco rigurosos junto a investigaciones bien diseñadas, lo que afecta la validez interna del procedimiento (Wolf, 1986). La mayoría de los expertos recomienda incluir todos los estudios independientemente de su calidad, codificándolos en función de la calidad del diseño empleado y examinando si los resultados difieren entre sí en función de ese criterio de calidad (Wolf, 1986; Gómez Benito, 1987; Rosenthal y Di Matteo, 2002). No obstante, en caso de optar por la exclusión de algún estudio, debe hacerse sobre la base de criterios explícitamente definidos e informados.

También se ha observado en este método el denominado problema de "las peras y manzanas". Esta metáfora se refiere a que las conclusiones derivadas de los meta-análisis algunas veces son inadecuadas ya que están basadas en la integración de

estudios que, por ejemplo, incluyen diferentes definiciones de variables y tipos de muestras o instrumentos. Si bien el investigador cuenta con procedimientos que permiten evaluar la homogeneidad de los resultados y los efectos mediadores, así como con diferentes métodos de ponderación de puntuaciones, debería ser siempre cuidadoso en la selección de los *inputs*

5
INTERPRETACIÓN DE PUNTUACIONES

Fabián Olaz - Silvia Tornimbeni

5.1. Interpretación referida a normas

5.1.1. Concepto

Un test está formado por varios ítems ante los cuales el individuo debe emitir sus respuestas. El resultado inicial de un test es el puntaje bruto, directo u original que se obtiene por la sumatoria de las respuestas correctas (en los tests de ejecución máxima) o respuestas clave (en los tests de comportamiento típico) (Walsh y Betz, 1990).

En el caso de los tests construidos sobre la base de la teoría de respuesta al ítem (TRI), esta puntuación original se denomina puntuación *theta*. A diferencia de las puntuaciones originales de tests basados en la teoría clásica de los tests (TCT), las puntuaciones theta no se obtienen de la simple sumatoria de las respuestas a los ítems, sino que resultan de la interacción de las respuestas del examinado con las características de los reactivos (la dificultad de los ítems, por ejemplo) (Hogan, 2004). Estas puntuaciones varían entre -4 y 4, aproximadamente, y se interpretan de manera semejante a las puntuaciones estándar que examinaremos más abajo.

Pese a las particularidades de cada teoría (TCT o TRI), las puntuaciones originales de tests son arbitrarias y, por lo general, no poseen un significado unívoco. De esta manera, por ejemplo, es poco esclarecedor conocer que un individuo resolvió 15 problemas en un test de habilidades matemáticas de 30 ítems. El hecho de que las puntuaciones originales no sean suficientes

para interpretar los resultados obedece a limitaciones de los tests como instrumentos de medición (Murat, 1985), tales como:

1. Carecen de cero absoluto (el cero en puntuación no indica ausencia absoluta del rasgo porque nunca se realiza un muestreo exhaustivo o representativo de sus posibles "indicadores operacionales").
2. No poseen unidades de medida constantes (situación muy diferente a otras unidades de medida, como el metro para medir la longitud o el kilogramo para el peso).

En los tests que miden rasgos latentes (como inteligencia o personalidad) la estrategia comúnmente empleada para atribuir significado a las puntuaciones originales es comparar los resultados individuales con las puntuaciones del grupo de referencia en la misma prueba.

De esta manera, los puntajes originales individuales son comparados con la distribución de puntajes de uno o más grupos de referencia (APA, 1999). Continuando con la ejemplificación anterior, si sabemos que el 60% de los estudiantes de un curso obtuvo puntajes de 15 o inferiores en el test de habilidades numéricas, hemos agregado significado al puntaje original de este estudiante.

Para poder realizar esta interpretación comparativa de los puntajes individuales se requiere un proceso denominado estandarización, el que se desarrolla mediante las siguientes operaciones:

a) Selección de una muestra representativa de la población meta para la cual se elaboró el instrumento, o muestra de estandarización.
b) Administración del test a esa muestra y registro de las puntuaciones originales de los individuos.
c) Transformación de las puntuaciones originarias en puntuaciones derivadas que indican la posición relativa de los puntajes directos individuales en relación con el grupo de referencia.

El producto final de este proceso de estandarización son los *baremos* de un test, que han sido definidos como tablas de equivalencia entre puntuaciones originarias y transformadas que permiten la comparación de los resultados individuales con los de un grupo de referencia (Grasso, 1999). El paso *b* del esquema anterior es relativamente rutinario y no presenta dificultades especiales para un profesional entrenado en el test en cuestión. Los pasos *a* y *c*, en cambio, requieren conocimientos específicos y se desarrollarán en el siguiente apartado.

Algunas precisiones semánticas previas son necesarias para evitar confusiones; en efecto, tal como define el diccionario de la Real Academia Española, los baremos son *normas establecidas por convención para evaluar los méritos personales*. Por consiguiente, ambos términos (normas y baremos) poseen el mismo significado en este contexto. Hemos optado, en general, por el término "baremo" para no generar confusión con las "normas" o estándares técnicos de los tests psicológicos, un concepto más general y anteriormente definido. No obstante, en la literatura psicométrica en español las dos palabras se usan indistintamente y con similar frecuencia. En inglés no existe esta dificultad, puesto que "*norms*" refiere a "baremos" así como "*standards*" a "normas técnicas".

5.1.2. Muestra de estandarización

Para Murat (1985), quien se propone aplicar un test tendrá que decidir entre: a) construir sus propias normas de interpretación de los puntajes (baremos), o bien, b) emplear los baremos elaborados por otro investigador.

En este último caso se deben tomar precauciones especiales antes de utilizar el test, tales como confirmar que los baremos estén actualizados y que la muestra de estandarización original sea semejante a la población meta de un test particular. Estas consideraciones adquieren especial relevancia en nuestro medio, dado que muchas veces no contamos con normas locales y el profesional se ve enfrentado a la difícil situación de escoger entre un baremo elaborado para una población diferente a la cual pertenecen los individuos que pretende evaluar o no hacer uso de baremo alguno.

Si se opta por el uso de baremos, la muestra de estandariza-

ción original debería ser lo más parecida posible a la población de aplicación actual del test en características demográficas tales como sexo, edad, nivel educativo y nivel socioeconómico u otras variables relacionadas con el desempeño en el test. Es muy importante que las muestras de estandarización sean cuidadosamente definidas y claramente descritas por los constructores del test para que, de esta forma, el usuario pueda escoger aquellos instrumentos cuyas normas sean apropiadas para su población meta.

No debe confundirse la elaboración de un baremo con el proceso más complejo y comprensivo de adaptación de tests, que no sólo implica poseer normas adecuadas sino replicar los estudios psicométricos esenciales del instrumento. La construcción de baremos es sólo una condición necesaria pero no suficiente para un empleo adecuado y éticamente responsable de tests elaborados en otros contextos socioculturales, problemática que se tratará en el capítulo final de este texto.

Para construir baremos se debe disponer de una muestra representativa de la población que será evaluada por medio de un test. Los baremos nacionales son extremadamente costosos y difíciles de obtener. La APA (1999) menciona diferentes tipos de normas (baremos) que pueden ser apropiadas para el uso de un test dado: normas locales (de un determinado lugar geográfico), normas regionales y normas específicas (por ejemplo, de una institución u ocupación). Es de especial importancia, sin embargo, que todos los baremos sean producto de un proceso de muestreo técnicamente riguroso.

Cuando un test es susceptible de aplicación a grupos distintos y existen diferencias significativas entre ellos en la variable medida por el test, deben elaborarse baremos separados de modo que cada persona pueda ser comparada con su verdadero grupo de referencia. Éste es, por ejemplo, el caso de los inventarios de intereses, en los que normalmente se presentan baremos diferenciados por sexo, puesto que las mujeres y los hombres (como grupo) difieren significativamente en sus perfiles de intereses vocacionales.

Existen voces críticas al empleo de baremos debido a las dificultades que se presentan para obtener muestras verdaderamente representativas de una población determinada y a la utilidad real de comparar los resultados individuales con un grupo

de referencia en situaciones de evaluación que no implican clasificación de personas, en especial cuando se emplean tests de comportamiento típico como los inventarios de intereses vocacionales o de personalidad (Goldberg, 1999; Cronbach, 1998). Ambas objeciones son atendibles y deberían considerarse antes del empleo o construcción de un baremo.

En particular, debe enfatizarse que si se construye un baremo para interpretar los puntajes de un test, debe ser generado a partir de una muestra representativa de la población meta de evaluación. De otro modo, la utilización del baremo puede conducir a intepretaciones equívocas y a errores considerables para el autoconocimiento de los individuos o la clasificación efectuada a partir de un test.

Sin recursos humanos y económicos considerables es muy dificultoso construir baremos nacionales o regionales. Una alternativa factible es la de obtener normas específicas (una institución, por ejemplo) o de utilidad local (una ciudad pequeña, por ejemplo) y para un grupo claramente definido (adolescentes urbanos que cursan el secundario básico, por ejemplo). A su vez, los usuarios deben estar atentos a las situaciones en las cuales los baremos son menos apropiados para algunos grupos de individuos que para otros. En un inventario de intereses ocupacionales, por ejemplo, los baremos utilizados para aquellas personas que se encuentran actualmente trabajando en alguna ocupación pueden ser inapropiados para interpretar los puntajes de los individuos desocupados (APA, 1999).

Las técnicas estadísticas para obtener muestras de estandarización van desde el muestreo aleatorio simple hasta estrategias más sofisticadas como el muestreo aleatorio estratificado, que reduce al mínimo la posibilidad de seleccionar una muestra no representativa (Aiken, 2003). En el muestreo aleatorio estratificado, la población meta es categorizada en una serie de variables (sexo, edad, nivel socioeconómico, lugar de residencia) que se supone poseen relación con el constructo medido por un test y luego se seleccionan aleatoriamente submuestras proporcionales de cada uno los estratos considerados.

Otro aspecto a tener en cuenta es el de actualización de los baremos. Algunos autores (Grasso, 1999; Aiken, 2003) recomiendan actualizar las normas cada cinco años, aproximada-

mente, e inclusive antes si se presenta un cambio significativo, tal como una modificación curricular importante en un nivel educativo. La actualización periódica de los baremos es un requisito básico para la validez de las interpretaciones de los puntajes de test referidos a normas (APA, 1999).

5.1.3. Métodos de transformación de puntuaciones

Recordemos que para interpretar comparativamente los resultados individuales de un test es necesario transformar las puntuaciones directas (originales) en otras derivadas. Las transformaciones para obtener estas puntuaciones derivadas pueden ser de dos clases: lineales y no lineales.

Transformaciones lineales

Una transformación es lineal cuando se obtiene una nueva escala de medición que respeta las distancias entre las unidades de medida de la escala original. Es decir, un cambio en la puntuación de la escala original se corresponde directamente con el cambio de puntuación en la escala transformada. De esta manera, la relación entre los intervalos es independiente de la unidad de medida empleada y del punto de origen de la escala. Esto se obtiene restando un valor constante de cada puntuación original y dividiendo el resultado por otra constante, como veremos más abajo.

Las transformaciones lineales no alteran la distribución original de frecuencias de las puntuaciones. Si la distribución original es normal continuará siéndolo después de ser transformada y si, por el contrario, presenta una asimetría positiva o negativa, estas características también se mantendrán luego de la transformación (Martínez Arias, 1995). Recuérdese que una distribución de frecuencias muestra la cantidad de sujetos que obtuvieron un valor determinado o están incluidos en una categoría de la variable medida. Cuando la información de una tabla de distribución de frecuencias se presenta gráficamente (histogramas o polígonos de frecuencias) hablamos de la forma de distribución de las frecuencias, tales como distribuciones normales o percentilares, entre otras.

Las transformaciones lineales más utilizadas son las puntuaciones estándar o *z*.

Puntuación estándar (*z*):

Como señala Aiken (2003), al transformar las puntuaciones originarias en puntajes *z* se obtiene una distribución que tiene la misma forma, pero una media y una desviación estándar diferentes a las de la distribución de la puntuaciones originarias. La media de las puntuaciones *z* es igual a 0 y la desviación estándar es igual a 1. Los puntajes equivalentes *z* de una distribución particular de puntuaciones originales pueden calcularse por medio de la siguiente fórmula:

$$zi = \frac{X \pm M}{s_x}$$

Esto es, la puntuación estándar de un sujeto es igual a su puntuación originaria (X) menos la media de las puntuaciones del grupo de referencia (M), dividido por la desviación estándar (s) (Murat, 1985). Al efectuar esta transformación, la media del grupo de referencia es el punto de origen de la nueva escala de medición y la unidad de medida será la desviación estándar. En otras palabras, las puntuaciones estándar expresan la distancia del individuo a la media en función de la desviación estándar de la distribución (Anastasi y Urbina, 1998).

Un ejemplo de cálculo de puntuaciones estándar con una media (M) de 60 y una desviación estándar (s) de 5 para dos individuos (S1 y S 2) sería el siguiente:

Puntuación S_1	*Puntuación S_2*
$x_1 = 65$	$x_2 = 58$
$z_1 = \frac{65 - 60}{5}$	$z_2 = \frac{58 - 60}{5}$
$z_1 = 1$	$z_2 = -0{,}40$

Tabla 5.1. Ejemplo de puntuaciones *z*.
Baremo del test de laberintos de Porteus

X	z
10,00	1,34
9,75	0,86
9,50	0,58
9,25	0,30
9,00	0,022
8,75	-0,25
8,50	-0,53
8,25	-0,82
8,00	-1,10

Las puntuaciones *z* pueden ser tanto negativas como positivas, y generalmente sus valores varían entre -3,00 y +3,00. Debido a las dificultades que ocasiona la presencia de valores negativos y decimales, se suele proceder a una segunda transformación lineal, en la que se multiplica cada puntuación *z* por una nueva desviación estándar fijada arbitrariamente por el examinador y se suma luego a ese resultado un valor establecido para la media. En este caso se tendrá una nueva distribución que conserva la forma de las puntuaciones originales, modificándose solamente la media y la desviación estándar, aunque las puntuaciones transformadas se expresan en una nueva escala. La fórmula para obtener esta segunda transformación es:

$$z' = z.k + m$$

Donde:

- z' = Puntuación transformada
- z = Puntuación estándar correspondiente a un individuo
- k = Desviación estándar establecida por el examinador
- m = Media constante establecida por el examinador

Hogan (2004) planteó una fórmula alternativa para el cálculo directo de puntuaciones *z'*, sin necesidad de realizar dos transformaciones sucesivas:

$$z' = \frac{S_e}{S_o}(X \pm M_o) + M_e$$

Donde:

- z' = Puntuación estándar que se desea obtener
- S_e = Desviación estándar establecida por el examinador
- S_o = Desviación estándar de los puntajes originales
- M_o = Media de los puntajes originales
- M_e = Media establecida por el examinador
- X = Puntaje original

Es importante señalar que, por lo general, la media y la desviación estándar son preestablecidas con la finalidad de facilitar la comparación con otros tests que miden un mismo constructo. De esta manera, los inventarios de personalidad utilizan frecuentemente una media de 50 y una desviación estándar de 10 (véase más adelante el cálculo de puntuaciones T), mientras que los tests de inteligencia usualmente emplean una media igual a 100 y una desviación estándar de 15.

El coeficiente de desviación es un caso especial de puntuación estándar, utilizado por las escalas de inteligencia de Wechsler (1999; 2005). En las escalas de inteligencia de Wechsler (WAIS, WISC) la media propuesta es 100 y la desviación estándar igual a 15. Un puntaje estándar de 100 en estos tests define el desempeño de un individuo de inteligencia promedio. Alrededor de los dos tercios de todos los individuos obtienen puntajes de entre 85 y 115 (que corresponden a una desviación estándar de 1, por encima y por debajo de la media, respectivamente), alrededor del 95% en el intervalo 70-130 (dos desviaciones estándar en ambas direcciones de la media) y casi todos obtienen puntajes entre 55 y 145 (tres desviaciones estándar a ambos lados de la media). La mayoría de los examinadores utilizan adicionalmente una notación cualitativa para describir la inteligencia de un individuo. De este modo, un coeficiente de desviación de 130 o superior se considera como muy superior al promedio, de 90 a 109 como equivalente al promedio y de 70 o menos como muy inferior.

Transformaciones no lineales

Las transformaciones no lineales, a diferencia de las lineales, asumen una distribución a priori (distribución normal, por ejemplo) que altera la forma de la distribución de los puntajes originales (Murat, 1985). Además, estas puntuaciones no pueden ser sumadas, promediadas o correlacionadas, puesto que no respetan las diferencias entre intervalos de la escala de medida original. Poseen la ventaja comparativa de resultar fácilmente comprensibles para personas sin conocimientos estadísticos especializados (maestros o estudiantes, por ejemplo). Las dos transformaciones no lineales más usuales son los percentiles y las puntuaciones estándar normalizadas.

a) Percentiles

Los percentiles expresan el porcentaje de personas, en un grupo de referencia, que queda por debajo de una puntuación original determinada. Así, por ejemplo, si el 30% de los individuos de una muestra de estandarización obtuvo un puntaje igual o inferior a 40 en un test, a una puntuación original de 40 le corresponderá un percentil 30 (P_{30}). Un percentil es un punto en la escala de medición originaria que divide el total de observaciones en dos partes. De este modo, el percentil 30 dejaría por debajo el 30% de los casos de la muestra de estandarización y por encima quedaría el 70% restante.

Con los percentiles empezamos a contar desde abajo, de tal forma que a un percentil más bajo corresponde una posición más baja del individuo en el test. El percentil 50 (P_{50}) corresponde a la mediana. El percentil 25 se corresponde con el primer cuartil ($Q_{1)}$ y el percentil 75 con el tercer cuartil ($Q_{3)}$. Estos dos últimos percentiles suelen utilizarse como puntos de corte para indicar la magnitud relativamente elevada y baja (respectivamente) de un atributo determinado.

Si bien el cálculo de percentiles se realiza fácilmente desde cualquier *software* estadístico (tal como SPSS, por ejemplo), a continuación se ejemplifica el cálculo de estas puntuaciones a los fines de esclarecer la lógica del procedimiento. La fórmula de cálculo para datos no agrupados es la siguiente:

$$P_x = \frac{(fa + 0.50.fp).100}{N}$$

Donde:

fa = Frecuencia acumulada hasta el puntaje original seleccionado
fp = Frecuencia propia del puntaje original seleccionado
N = Número total de casos

Para ejemplificar el calculo de percentiles utilizando esta formula se utilizarán los datos de la tabla siguiente.

Tabla 5.2. Datos sin agrupar del CIP-R

Puntaje original	Frecuencia	Frecuencia acumulada
27	32	809
26	25	777
25	33	752
24	24	719
23	24	695
22	25	671
21	27	646
20	31	619
19	31	588
18	42	557
17	35	515
16	50	480
15	54	430
14	38	376
13	52	338
12	62	286
11	73	224
10	74	151
9	77	77

Si, por ejemplo, quisiéramos calcular el percentil para la puntuación original 25, tendríamos:

$$P_x = \frac{(719+0{,}50.\ 33)\ .\ 100}{809}$$

$$P_x = \frac{(719 +16{,}5)\ .100}{809} = 90{,}91$$

Es decir que a una puntuación bruta de 25 le corresponde un percentil equivalente a 90,91.

El sumar a la frecuencia acumulada la mitad de la frecuencia correspondiente a la puntuación para la cual queremos calcular el percentil (0,50 x *fp*), en este caso la puntuación 25, se debe a que se supone que la puntuación 25 representa un intervalo que va desde 24,5 a 25,5, en el que se reparten de forma homogénea todas las frecuencias. El punto medio de este intervalo es 25, por lo que se le asignan hasta ese punto la mitad de las frecuencias (Martínez Arias, 1995).

La tabla 5.3. presenta un baremo expresado en percentiles.

La facilidad de interpretación de los percentiles los hace especialmente atractivos. Pero, a pesar de su sencillez, tienen una desventaja considerable si se los compara con las transformaciones lineales. Al respecto, debe recordarse que los percentiles operan en un nivel de medición ordinal y no intervalar, como los puntajes *z*. Por consiguiente, este tipo de puntuaciones derivadas altera profundamente la distribución de las puntuaciones originales, transformándola en una nueva distribución con marcada desigualdad de las unidades en diversos puntos de la escala (Hogan, 2004).

Este problema no es trivial puesto que una diferencia en el puntaje original del test implicará muchos percentiles de distancia a la mitad de la distribución, debido a que los percentiles tienden a agruparse en el medio de la distribución, pero sólo una diferencia mínima en percentiles en los extremos de la distribución (Aiken, 2003). Así, por ejemplo, en el inventario NEO-PI-R (Costa y McCrae, 1999), a la distancia entre una puntuación original de 27 y 30 (3 puntos) le corresponde una diferencia

Tabla 5.3. Baremo en percentiles del CIP-R

Percentil	B	C	D	E	F	G	H	I	J	K	L	M	M
1	6	9	8	10	8	8	8	8	8	7	6	8	6
5	6	9	8	10	8	8	9	9	8	7	6	9	6
10	6	10	8	11	8	8	11	10	9	7	7	11	7
25	8	11	9	12	8	9-	13	12	10	7	8	13	8
40	10	12	11	15	9	11	16	14	11	8	10	15	9
50	12	14	13	17	10	13	18	16	13	9	12	17	11
60	13	16	16	19	10	14	19	17	15	11	13	18	12
70	14	18	18	21	12	15	21	19	17	13	15	20	13
75	15	19	19	22	12	16	21	20	18	14	16	20	14
80	16	21	20	24	13	17	22	21	19	15	16	21	15
90	17	24	22	27	15	19	23	23	21	18	17	23	16
95	18	26	24	29	17	22	24	24	23	20	18	24	18
99	18	27	24	30	20	24	24	24	24	21	18	24	18

Ref.: A = Lingüística, B = Musical, C = Humanística, D = Económica, E = Tecnológica, F = Naturalista, G = Asistencial, H = Artística, I = Sanitaria, J = Cálculo, K = Jurídica, L = Comunicacional, M = Científica.

de seis unidades percentilares (percentiles 93 y 99); en cambio, en el medio de la distribución, a una distancia de 3 unidades en puntuación original (20-23, por ejemplo) le corresponde una diferencia de 27 unidades en percentiles (43-70). Esto puede ser particularmente problemático cuando se utilizan puntuaciones percentilares en contextos de clasificación (véase el capítulo 4 sobre validez) o selección de personas.

b) Puntuación estándar normalizada (zn)

Para facilitar la comparación de diferentes puntuaciones transformadas (puntajes *z* con percentiles, por ejemplo), se suele recurrir a transformaciones no lineales que modifican la forma de distribución de las puntuaciones originales, convirtiéndolas en una distribución normal.

En psicometría es muy importante la distribución normal, un modelo estadístico que permite estimar probabilidades de ocu-

rrencia de los diferentes valores de una variable pero que no se corresponde exactamente con ninguna forma de distribución de frecuencias real u observada (Grasso, 1999). Una distribución normal se representa gráficamente por medio de la curva normal, en forma de campana. La curva normal posee propiedades matemáticas de gran importancia y sirve de fundamento a varios tipos de análisis estadísticos. La curva es simétrica bilateralmente con un punto máximo hacia el centro de la distribución e indica, esencialmente, que el mayor número de casos se agrupa hacia el centro, disminuyendo gradualmente en ambas direcciones a medida que nos alejamos del centro de la distribución. La mayoría de los atributos psicológicos, evaluados a través de tests, poseen una distribución aproximada a la normal (Cohen y Swerdlik, 2000).

El procedimiento empleado en este tipo de transformaciones no lineales se denomina *normalización* y las puntuaciones obtenidas mediante este proceso reciben el nombre de *puntuación estándar normalizada*. Algunos casos especiales de este tipo de puntuaciones son las denominadas "T" (en honor a Terman) con media igual a 50 y desviación estándar igual a 10, y las puntuaciones estaninas con media = 5 y una desviación estándar = 2. El nombre *estanina* (contracción de *standard nine*) se basa en que las unidades de estas puntuaciones transformadas van de 1 a 9. Una dificultad de las estaninas es que el uso de un sólo dígito puede sugerir diferencias significativas entre dos individuos cuando éstas no son tales (Hood y Johnson, 2002).

La transformación de puntuaciones originales a puntuaciones estándar normalizadas se realiza mediante el siguiente procedimiento:

1. Estimar el percentil correspondiente a una puntuación original.
2. Convertir ese percentil en una proporción.
3. En el cuadro de áreas por debajo de la curva normal, ubicar la puntuación *z* debajo de la cual se encuentra esa proporción. Por ejemplo, dada una proporción de 0,97, la puntuación z correspondiente es de 1,89.
4. Proceder al cálculo de la puntuación T u otra semejante mediante la ecuación:

$$z' = z.k + m$$

Donde:

z' = Puntuación estándar normalizada
z = Puntuación estándar correspondiente a un puntaje bruto determinado
k = Desviación estándar (en el caso de los puntajes T igual a 10)
m = Media (50 para las puntuaciones T)

Tabla 5.4. Intervalos de puntajes originales (X), percentil correspondiente (Pc), *z* normalizado (*zn*) y puntaje T correspondiente

X	Pc	zn	T
119-192	99	2,33	73
113-118	98	2,06	71
108-112	97	1,89	69
105-107	96	1,76	68
99-104	95	1,65	66
92-98	90	1,29	63
87-91	85	1,04	60
82-86	80	0,85	58

Por ejemplo, a un puntaje original en el intervalo 108-112 del cuadro anterior, le corresponde un percentil 97. Transformando ese percentil en proporción tenemos un valor de 0,97. En el cuadro de áreas bajo la curva normal, a una proporción de 0,97 le corresponde una puntuación *z* de 1,8. La puntuación original 109 (equivalente a un percentil 97), por ejemplo, supera entonces al 97% de los casos y se encuentra a 1,8 desviación estándar por encima de la media de las puntuaciones.

Las áreas de la distribución normal pueden consultarse en la mayoría de los textos de estadística e indican la proporción de casos correspondientes a una puntuación *z* determinada. En el

texto de Martínez Arias (1995) se presentan tablas exhaustivas de las áreas de la distribución normal, comprendidas entre *z* -3 a 3.

Aplicando a los datos del ejemplo anterior la fórmula de la transformación lineal, se tiene que:

$$T = 1.\ 89\ .\ 10 + 50 = 68.9$$

Redondeando este valor final tenemos un T = 69. Como puede observarse en la tabla precedente, la puntuación T correspondiente al percentil 97 es 69.

Las puntuaciones T se distribuyen en un rango que va desde 20 (aproximadamente 3 desviaciones estándar por debajo de la media) a 80 (3 desviaciones estándar por encima de la media). No deben confundirse estas puntuaciones con los valores *t* de Student utilizados en las pruebas estadísticas de significación. La decisión de normalizar las puntuaciones no debería tomarse sin cuidado; por ejemplo, no es recomendable cuando la distribución de puntuaciones originales del test se aleja considerablemente de una distribución normal (Martínez Arias, 1995).

Las puntuaciones T también pueden obtenerse de manera más directa, utilizando la fórmula cálculo de *z'*, en este caso con una media de 50 y una desviación estándar de 10. Con este procedimiento no se altera la forma de la distribución de los puntajes originales como acontece cuando se utiliza el procedimiento de normalización anteriormente descrito, y esta variedad de puntajes T se convierte en otro caso de transformación lineal, que revisamos en el apartado anterior (Kaplan y Saccuzzo, 2006).

En la actualidad, todos los cálculos requeridos para construir baremos se realizan por medio de programas estadísticos computarizados. Sin embargo, conocer algunos procedimientos básicos de cálculo como los anteriores facilita una mejor comprensión de la lógica de los mismos así como una interpretación adecuada de las salidas (*outputs*) de la computadora.

5.2. Otros métodos de interpretación de puntuaciones

5.2.1. Puntuaciones ipsativas

Si bien la interpretación referida a normas es la más utilizada, existen otras formas de interpretación de los puntajes originales de un test. Una de estas formas alternativas de interpretación son las puntuaciones ipsativas. Estos puntajes se obtienen en tests que utilizan un formato de ítems de elección forzada, donde el examinado debe optar por una alternativa entre varias que lo describen. En estos tests, los ítems se califican de tal manera que la elección de una de las opciones de respuesta produce un incremento en la puntuación de una escala o dimensión medida y al mismo tiempo una disminución en el puntaje de otra de las escalas o dimensiones del test. Por consiguiente, este tipo de puntuaciones muestra la fuerza "relativa" de las puntuaciones en lugar de la fuerza "absoluta" de las mismas (Hogan, 2004). Para comprender cabalmente esta última afirmación considérese el siguiente ejemplo.

A los fines de medir intereses vocacionales se pueden utilizar dos formatos de respuesta diferentes:

Formato A:

Seleccione de cada par de actividades aquella que más le interese:

a. Resolver ecuaciones matemáticas o **b**. Aprender estilos de pintura artística

a. Tocar un instrumento musical o **b**. Hacer cálculos numéricos

Formato B:

Examine cada ítem e indique con una cruz en el casillero correspondiente su Desagrado (D), Indiferencia (I), o Agrado (A) por el mismo.

	D	I	A
1. Resolver ecuaciones matemáticas			
2. Aprender estilos de pintura artística			
3. Hacer cálculos numéricos			
4. Tocar un instrumento musical			

En el formato B, si asignáramos un puntaje 1 a la opción **D** (Desagrado), un 2 a **I** (Indiferencia) y un 3 a la opción **A** (Agrado), la persona puede elegir cualquier puntuación para cada uno de los ítems, siendo independiente entre sí la puntuación asignada a cada uno de ellos.

El formato A es característico de una puntuación ipsativa. Como puede apreciarse, los ítems de este tipo de tests exigen una elección entre ambas opciones de respuesta, representadas respectivamente por *a* y *b*, y no es posible elegir o rechazar ambas. Además, en la medida en que aumenta la puntuación en la escala "artística", por ejemplo, disminuye o no aumenta su puntuación en "cálculo", evidenciando la fuerza relativa de los intereses. Es probable que a una persona le desagraden o agraden ambas actividades aunque prefiera (o rechace) una de ellas más que la otra (Hogan, 2004).

Como puede apreciarse, las puntuaciones finales de una escala ipsativa expresan un perfil de los "puntos" fuertes y débiles de un individuo sin compararlo con un grupo de referencia. Con este procedimiento, la deseabilidad social (un ítem es escogido sólo porque expresa una idea socialmente aceptable) y los sesgos individuales de respuesta (los individuos eligen siempre la misma opción de respuesta o emiten respuestas similares a cada ítem) se controlan exitosamente.

La interpretación ipsativa posee la limitación de obstaculizar la aplicación de algunos estadísticos usuales en psicometría debido a la falta de independencia de sus ítems (Kerlinger y Lee, 2002). Otra dificultad inherente a estas puntuaciones es la resistencia que despiertan los ítems de elección forzada en muchos individuos. Algunos tests muy populares en contextos de orientación que utilizaban solamente puntuaciones ipsativas o puntajes originales (Registro de Preferencias Kuder y Self-Directed Search, respectivamente) incluyen, en sus últimas versiones, baremos para interpretar los resultados.

Otra excepción a la interpretación referida a normas son los tests referidos a criterio o dominio, que se expondrán a continuación.

5.2.2. Interpretación referida a criterio

Tal como se explicó más arriba, puesto que las puntuaciones originales de un test carecen de un significado unívoco, normalmente se interpretan comparándolas con un grupo de referencia. Una forma alternativa de interpretación de las puntuaciones consiste en compararlas con un criterio de logro u objetivo a alcanzar, previamente especificado.

Por ejemplo, si consideramos que un criterio de desempeño es responder en forma correcta al menos 14 de una serie de 20 preguntas en una prueba de conocimiento de literatura, 14 sería el estándar con el cual comparar los aciertos obtenidos por un individuo en ese test de rendimiento. En otro ejemplo, no sería relevante si se ubica en el percentil 90 en un examen de conducción de vehículos, por ejemplo. Si alguien consistentemente no respeta el semáforo, no es un buen candidato para recibir una certificación de maestría de ese dominio, aunque su puntuación en el test haya sido elevada en relación con el grupo de referencia que respondió el mismo test (Woolfolk, 2006).

Este tipo de interpretación de puntuaciones se denomina "interpretación referida a criterio o a dominio" y los tests que la utilizan, "tests con referencia a criterio o dominio". Estos tests presuponen que existe un área específica o dominio de conocimiento o habilidad que puede ser claramente definido y delimitado.

Una prueba referida a criterio es aquella que deliberadamente se construye para conducir a medidas directamente interpretables en términos de pautas específicas de desempeño (Glaser, 1963), las cuales se determinan definiendo una clase o dominio de tareas que el individuo debe realizar. Popham (1975) afirmó que los tests referidos a criterio se utilizan para evaluar la posición absoluta de un individuo con respecto a algún dominio de conductas previamente definido. Es importante considerar que en muchos dominios es dificultoso fijar con precisión objetivos específicos y, además, en varias ocasiones el establecimiento de un criterio de desempeño (25 respuestas correctas y no 24, por ejemplo) es bastante arbitrario.

Los tests referidos a criterio se desarrollaron a comienzos del siglo XX. Sin embargo, esta línea de investigación fue abandonada durante el período comprendido entre las dos guerras

mundiales (1914-1945), para luego ser retomada a mediados del siglo pasado (Martínez Arias, 1995). Recién en los años setenta se comenzó a aplicar sistemáticamente este tipo de pruebas, en particular en la evaluación educativa.

La evaluación en educación se realiza con diferentes fines, entre ellos:

a) determinar la calidad de un sistema educativo
b) evaluar la adecuación de un currículo
c) evaluar los efectos de un programa de enseñanza
d) evaluar el rendimiento de los estudiantes
e) seleccionar aspirantes a un curso o carrera

La evaluación educativa comenzó como un medio para seleccionar alumnos, y los tests que más se utilizaron fueron los referidos a normas, cuyos resultados se interpretan en función de un grupo normativo o baremo. Siguiendo este modelo, posteriormente se construyeron tests referidos a normas para ser aplicados con otros propósitos, tales como evaluar la calidad de un sistema educativo o el rendimiento académico. No obstante, algunos especialistas en educación advirtieron que la aplicación de este tipo de pruebas no proporcionaba información adecuada si, por ejemplo, el propósito esencial era evaluar el logro de los objetivos propuestos por un sistema educativo.

En los años sesenta, junto con la instrucción programada y otros programas educativos semejantes, surge la necesidad de una evaluación diagnóstica previa de los individuos y, a posteriori, para verificar los cambios en los mismos como efecto de la aplicación de esos programas. Estas razones impulsaron el desarrollo de este enfoque alternativo en la interpretación de puntuaciones de tests, donde no interesa tanto comparar al individuo con la población a la cual pertenece, sino medir cambios de cada individuo a lo largo del aprendizaje.

La evaluación referida a criterio supone una filosofía diferente del quehacer educativo. Tal como argumenta Tyler (1978), la función esencial del maestro no es identificar a los mejores y a los peores alumnos sino tratar de que todos los estudiantes logren los objetivos relacionados con dominios de aprendizaje específicos.

Según Bond (1996) el contenido de un test referido a normas

es seleccionado con el fin de obtener información que permita discriminar entre estudiantes, mientras que en las pruebas referidas a criterios, el contenido se selecciona sobre la base de su importancia para el currículo. Los tests con referencia a normas se basan en las diferencias individuales y, por lo tanto, tienen como objetivo primordial la selección y la predicción (Glaser, 1963). Por el contrario, las pruebas con referencia a criterio intentan medir cambios en los propios individuos o grupos como efecto de una intervención educativa. Por consiguiente, estos tests resultan más adecuados para fines de diagnóstico y prescripción de las experiencias de aprendizaje requeridas para asegurar el logro de determinados objetivos.

Como afirma Hogan (2004), es más apropiado hablar de interpretación referida a normas y a criterio, puesto que la puntuación de un mismo test puede interpretarse de estas dos formas. Imaginemos un test de aritmética elemental, con ítems relacionados con operaciones básicas (suma, multiplicación, división, resta) para estudiantes de cuarto grado. Podría fijarse un punto de corte del 75% de los ítems acertados (25 ítems, por ejemplo) como indicador de rendimiento satisfactorio en ese dominio (aritmética elemental). Ésta es una interpretación relacionada con criterio. Por el contrario, las puntuaciones del mismo test podrían interpretarse en referencia a normas si esa puntuación de 25 (los ítems acertados) se comparase con el rendimiento de la población meta (por ejemplo, todos los estudiantes de cuarto grado de una ciudad) y se determinara que es equivalente al percentil 75 de la muestra de estandarización del test, por ejemplo.

Los tests referidos a normas sugieren más bien cuánto han aprendido los individuos pero no esclarecen adecuadamente qué han aprendido. En cambio, las pruebas con referencia a criterio informan la posición absoluta de un sujeto en relación con un dominio conductual definido explícitamente. De este modo, cualquier cambio de posición del individuo en ese dominio adquiere un significado más claro, pues refleja un cambio interpretable en términos conductuales (Himmel, 1979). Cada estudiante es capaz o no de exhibir una habilidad particular, generar un producto específico o manifestar cierta conducta (Popham, 1975), y su desempeño debería valorarse con respecto a qué alcanzó co-

mo logro, y no con referencia a cuánto logró en comparación en sus compañeros.

La interpretación referida a criterio es aplicable sólo en dominios específicos de contenido, tales como aritmética, ortografía o las habilidades requeridas para el ejercicio de una ocupación (Hogan, 2004). El análisis de los resultados en este tipo de pruebas puede realizarse distinguiendo las habilidades o conductas en relación con un contenido temático que presentan mayor dificultad y las que son más fáciles de adquirir. Para ello, puede computarse el número de estudiantes que dan la respuesta correcta en cada ítem y dividirse esa frecuencia por el número total de estudiantes.

Como ejemplo, se presentan e interpretan los resultados obtenidos en una prueba de Lengua aplicada a niños de 6º grado de la Provincia de Córdoba (Ferreyra, 1982) con el propósito de evaluar el rendimiento de los alumnos al terminar el ciclo primario. En este caso, se fijaron los siguientes criterios: si el porcentaje de aciertos en el ítem era igual o menor al 30% se consideraron difíciles, si era igual o mayor al 70% se consideraban fáciles. Sobre la base de esos porcentajes, se examinaron los

Tabla 5.5. Resultados de una prueba de Lenguaje**.** Objetivo de conocimiento de criterios

ÍTEM	Total	Mujeres	Varones	ED	PD	EN
1. Identificar el tipo de lenguaje en un texto discursivo	0,593	0,603	0,585	0,570	0,638	0,531
5. Reconocer una noticia en artículos periodísticos	0,255	0,236	0,272	0,261	0,244	0,219
18. Identificar fuentes de información	0,868	0,902	0,840	0,855	0,897	0,750
26. Identificar el carácter de un texto descriptivo	0,584	0,630	0,544	0,580	0,592	0,531
46. Reconocer un género literario	0,263	0,296	0,235	0,254	0,282	0,219

Nota: ED: estatal diurna; PD: Privada diurna; EN: estatal nocturna.

ítems correspondientes a cada objetivo para estimar el grado de logro del mismo. Con este mismo criterio se analizaron los porcentajes de respuestas discriminadas según las variables consideradas: turno de asistencia (escuelas diurnas y nocturnas), sexo y tipo de escuela (estatal o privada).

De esta manera se identificaron, por ejemplo, diferencias en el rendimiento de los alumnos de escuelas diurnas y nocturnas en relación con la habilidad de interpretación. A continuación se analizan los resultados obtenidos para las preguntas correspondientes al objetivo "conocimiento de criterios".

La inspección de los datos anteriores permite inferir que, en general, los alumnos no logran reconocer una noticia entre los textos periodísticos o distinguir los géneros literarios (los porcentajes de acierto oscilan entre el 25 y el 29%). En cambio, sí conocen aquellos criterios que les permiten identificar distintas fuentes de información (los porcentajes de acierto oscilan entre el 75 y el 90%). Asimismo, se observa menor rendimiento en los alumnos de escuelas nocturnas que en los de escuelas diurnas. Entre estas últimas, las escuelas privadas obtienen porcentajes de acierto ligeramente superiores que las oficiales, y lo mismo se observa entre varones y mujeres (ligeramente superior en las mujeres). En síntesis, y en referencia al objetivo "conocimientos de criterios", se puede concluir que el mismo no es logrado en todos sus niveles o dimensiones por los estudiantes de esta muestra.

6
CONSTRUCCIÓN DE TESTS

Edgardo Pérez - Silvia Tornimbeni

Nuestra exposición se concentrará en los métodos de construcción de tests propuestos por la teoría clásica de los tests (TCT). Hemos elegido este aspecto debido al carácter introductorio de este texto y a que la mayoría de los tests publicados continúan elaborándose en el marco de esa teoría. No obstante, es creciente la influencia de la teoría de respuesta al ítem (TRI) con sus distintos modelos, por lo cual también se comentarán algunos procedimientos de construcción relacionados con la TRI.

Adicionalmente debe considerarse que los diversos tipos de tests (ejecución máxima, escalas de actitudes, referidos a criterio, etc.) utilizan procedimientos de construcción que difieren en algunos aspectos. Aquí intentaremos destacar aquello que los tests tienen en común en su metodología de construcción, aunque se revisarán algunas características particulares de construcción de los tests referidos a criterio o dominio. En general, el procedimiento estándar que se usa para construir cualquier test comprende los siguientes pasos (Herrera Rojas, 1998):

a. Delimitación del dominio del test, características de la población a la cual va dirigido y estructura formal del test (instrucciones, contenido y formato de respuesta a los ítems).
b. Redacción de los ítems.
c. Revisión de los ítems por expertos.
d. Análisis de las propiedades psicométricas de los ítems y/o escalas del test.
e. Elaboración de los materiales definitivos de prueba (manual, cuadernillo de ítems, hojas de respuesta).

Seguidamente analizaremos esta serie de pautas fundamentales para la construcción de un test.

6.1. Definición del dominio

La construcción de un test requiere, en primer lugar, un exhaustivo análisis conceptual del dominio o constructo a medir. Este análisis implica la selección y revisión de las teorías más relevantes, rigurosas y contemporáneas en relación con el constructo que se pretende medir. Se deben obtener definiciones conceptuales ajustadas del constructo o dominio de interés, así como seleccionar los indicadores operacionales adecuados para medirlo.

Todas las dimensiones importantes del constructo deben incluirse; si por ejemplo se quisiera medir "habilidades para el estudio", el test debería contemplar todas las sub-habilidades implícitas en ese dominio, tales como manejo y planificación del tiempo de estudio, elaboración de resúmenes, preparación de exámenes y toma de apuntes.

Bandura (2001) proporcionó un ejemplo de construcción de una escala de autoeficacia para el manejo del peso corporal. Puesto que el peso depende de factores tales como las calorías ingeridas, el nivel de ejercicio para quemar esas calorías y de factores genéticos que regulan los procesos metabólicos, la conducta de autocontrol del peso será mejor predicha por una escala que incluya ítems que contemplen equitativamente los factores causales y no se limite, por ejemplo, a evaluar solamente los hábitos alimenticios y la ingesta de calorías.

Si se trata de un test construido para medir rendimiento en un dominio específico, la definición de este último se realiza delimitando el universo de situaciones a ser evaluadas. Así, por ejemplo, en el caso de una prueba construida para evaluar conocimientos de estadística se deberían contemplar los objetivos, actividades y contenidos del programa de esa asignatura.

Como ya se mencionó, en la medición del rendimiento se pueden utilizar tests referidos a criterio o referidos a normas. Los procedimientos de construcción de las pruebas con referencia a criterio (o dominio) difieren de aquellos usados en las pruebas de rendimiento.

Para la elaboración de pruebas de rendimiento referidas a normas se parte de la construcción de una tabla de especificaciones. Ésta consiste en una tabla de doble entrada por medio de la cual se relacionan los objetivos cuyo logro se desea evaluar con los contenidos específicos correspondientes. Tomando esta tabla como marco de referencia se determina la cantidad de ítems que conformarán la prueba y se redactan los mismos.

En la construcción de una prueba con referencia a criterio, en vez de elaborar una tabla de especificaciones, se define y delimita el dominio de comportamientos correspondientes a cada objetivo. Al elaborar este tipo de instrumentos, un requisito fundamental es definir con claridad las habilidades o conocimientos que la prueba intenta evaluar. El dominio seleccionado (aritmética, por ejemplo) debe subdividirse en unidades pequeñas definidas en términos de ejecución, por ejemplo, "dividir números de tres dígitos por otros de dos dígitos". Después de que se han definido estos objetivos de aprendizaje se deben elaborar ítems para evaluar cada uno de ellos (Anastasi y Urbina, 1998). Según Hambleton y Rogers (1991), el "dominio" puede ser de conductas, objetivos y competencias, y su amplitud varía en función de la finalidad del test. Si el dominio comprende más de un objetivo, pueden construirse subtests para cada objetivo, y se evalúa el rendimiento de los sujetos en cada uno de ellos.

Existen varios procedimientos recomendados para la especificación del dominio de conductas o clase de tareas que el individuo debe realizar. En general siguen el siguiente esquema:

a. Definición del objetivo: se establece cuál o cuáles serán los objetivos que se evaluarán a través de la prueba, por ejemplo la habilidad de comprensión lectora, que incluye aquellas conductas o respuestas que se refieren únicamente a una comprensión de los mensajes literales contenidos en una comunicación textual.
b. Indicadores operacionales del objetivo: se describen, ahora en términos de conductas observables, el/los objetivos a ser evaluado/s. Siguiendo con el ejemplo anterior, un indicador operacional de la habilidad de comprensión podría ser "resumir adecuadamente un texto breve".

c. Especificación de las características de la situación de evaluación: por ejemplo, en un texto de divulgación científica, seleccionar las ideas principales y parafrasear el contenido de las mismas.
d. Características de la respuesta: se especifica cuál es la respuesta que se espera del estudiante evaluado, por ejemplo, que seleccione correctamente las ideas principales.

Además de definir el dominio es necesario delimitar aspectos complementarios del test, tales como la finalidad y la población meta del test (por ejemplo, un inventario de autoinforme para evaluar el autoconcepto en niños), el modo de aplicación (individual o colectivo, por ejemplo), el formato de respuesta (dicotómica o tipo *likert*, por ejemplo) y el tiempo de administración (duración del test), entre otras consideraciones preliminares (Hogan, 2004). El plan inicial del test también debe prever las instrucciones de administración y el modo de calificación e interpretación de las respuestas (puntuaciones originales, transformadas o ipsativas, por ejemplo).

6.2. Redacción de los ítems

Existen pautas convencionales para la redacción de ítems de tests. Éstas incluyen recomendaciones del tipo:

- Redactar ítems congruentes con el objetivo de medición.
- Evitar los ítems demasiados largos (de más de 20 vocablos).
- Evitar las oraciones complejas con ambigüedades de sentido.
- Evitar las frases con doble negación.
- Evitar el uso de expresiones extremas (nunca, siempre, todos).
- Utilizar el lenguaje más apropiado al nivel de maduración y educativo de la población meta de la medición (Oesterlind, 1990).

Recientemente, Moreno, Martínez y Muñiz (2004) han formulado otras directrices útiles para la redacción de ítems de elección múltiple (*multiple choice*) que comentaremos en esta sección.

Para Nunnally (1991), los dos errores más comunes en la redacción de ítems son: a) la ambigüedad, con preguntas vagas que admiten varias respuestas, por ejemplo, "¿que pasó con el Arte en el siglo XV?", y b) la trivialidad, al centrarse en aspectos poco importantes del constructo o dominio, por ejemplo, requerir la memorización de fechas irrelevantes. Bandura (2001) recomienda adicionalmente evitar el argot técnico que no forma parte del lenguaje cotidiano y los ítems que incluyen aspectos diferentes (multidimensionales) de un constructo para los cuales los individuos pueden tener diferentes percepciones, tales como: ¿cuán seguro te sentís de nadar y remar adecuadamente? Es obvio, en el ítem anterior, que una persona puede sentirse competente para nadar pero no para remar, y viceversa.

En la evaluación educativa, merece un apartado especial la construcción de pruebas objetivas con preguntas cerradas, ya sea del tipo verdadero/falso o de alternativas múltiples. Según Bloom (1966), estas pruebas son útiles para la medición de algunos objetivos cognoscitivos de nivel básico, tales como:

- *Recordar* (creador del coeficiente de correlación, por ejemplo).
- *Comprender* (el concepto de confiabilidad, por ejemplo).
- *Aplicar* un concepto general o utilizar información para resolver un problema (dada la media y la desviación estándar de una distribución, obtener la puntuación estándar correspondiente al puntaje original X).
- *Analizar,* que se refiere al pensamiento crítico, es decir, a identificar causas y realizar inferencias en base a información específica (interpretar los coeficientes alfa del test X e indicar qué factores pueden haber afectado la consistencia interna de ese test).

Para los objetivos cognoscitivos de nivel superior, tales como evaluar (juzgar el valor de materiales, tests o métodos estadísticos, por ejemplo) y crear (diseñar una investigación para verificar la estabilidad de un test, por ejemplo), se requiere otro tipo de pruebas, tales como las de preguntas abiertas o ensayo, así

como ítems que combinan la computación con el audio, el video y la realidad virtual en la formulación de las preguntas y el formato de respuesta, dentro de la denominada evaluación auténtica (Moreno, Martínez y Muñiz, 2004). Seguramente la evaluación del futuro exigirá pruebas con ítems que permitan medir de manera más pertinente el pensamiento creativo (divergente) y la resolución de problemas reales de una disciplina (Woolfolk, 2006).

A continuación se explicitan algunas recomendaciones para la construcción de ítems en las pruebas de opciones múltiples, puesto que son difíciles de elaborar adecuadamente. En ese sentido e ingeniosamente, Woolfolk (2006) comentó que muchos estudiantes llaman a estas pruebas "de adivinación múltiple", por lo mal que frecuentemente se elaboran.

Estos tests incluyen un enunciado, tronco o base, por ejemplo: "el método más adecuado para evaluar la estabilidad temporal es…" y una serie de alternativas o respuestas posibles, tales como: a) partición en mitades, b) acuerdo de jueces, y c) test-retest.

Con referencia al enunciado o base del ítem, las principales recomendaciones son:

1. Debe contener un esquema de indagación completa (que el estudiante no necesite leer las alternativas para emitir la respuesta correcta).
2. Se debe incluir lo estrictamente necesario para la comprensión de las respuestas. *Una ventaja de las puntuaciones estándar es…*, por ejemplo, y no: *Hay varios tipos de puntuaciones derivadas. La puntuación estándar es especialmente ventajosa por…*
3. Es preferible que las palabras que puedan repetirse en las alternativas se incluyan sólo en la proposición base. Un ítem del tipo de: "Una puntuación percentil: a) *indica* el porcentaje de ítems que se respondieron de manera correcta; b) *indica* el porcentaje de casos que obtuvieron una puntuación igual o menor a cierta puntuación original", etc*.,* por ejemplo, puede mejorarse con una base que exprese: "Una puntuación percentil *indica"*, evitando repetir "indica" en las alternativas.

4. Se deberá evitar redactar la proposición base como enunciado negativo, a menos que la finalidad sea reforzar el aprendizaje de lo que no debe hacerse.
5. La base no debe contener expresiones que puedan debilitar o confundir la respuesta correcta.
6. Cuando se intenta evaluar la comprensión de términos, es preferible que estos conceptos se mencionen en la base, y las descripciones o definiciones se incluyan en las alternativas de respuesta.
7. Debe evitarse que el ítem se refiera a contenidos triviales. Lo esencial del contenido debe incluirse en la base, no en los distractores, para evitar la lectura de material extenso o redundante que dificulte la comprensión del ítem.

Con referencia a las alternativas de respuesta (distractores y clave u opción correcta):

1. El ítem deberá contener una sola opción correcta, la cual tiene que estar acompañada por distractores que sean plausibles para el estudiante que no conoce la respuesta correcta y fácilmente desechables para el que la conoce.
2. Todas las alternativas deberán ser gramaticalmente semejantes e igualmente aceptables desde el sentido común. La distancia conceptual entre la opción correcta y los distractores debe ser amplia, pero lo suficientemente limitada como para que no se rechace a estos últimos por obvios.
3. Por lo general, tres alternativas de respuesta son suficientes puesto que el formato de cuatro opciones es más dificultoso de elaborar y, muchas veces, la elección de la última opción de respuesta resulta algo forzada. Redactar tres alternativas para un contenido determinado es más sencillo e igualmente confiable.
4. En cuanto al formato, se deberá evitar que la alternativa correcta sea la más larga.
5. Se deberán evitar las expresiones muy literales que expliquen el texto de estudio y que favorezcan la mera memorización.
6. Las alternativas incorrectas deberán tener el mismo grado de especificidad que la opción correcta de respuesta.

7. La alternativa correcta deberá estar dispuesta aleatoriamente. En el conjunto de ítems que componen una prueba, la opción correcta debe estar repartida entre las distintas ubicaciones posibles (a, b, y c, por ejemplo).
8. Debe evitarse que un ítem pueda ayudar a la respuesta correcta de otro.
9. Las distintas opciones de respuesta al ítem tienen que ser independientes entre sí, sin solaparse y sin referirse unas a otras pues ello introduce dificultades o facilidades indebidas. Por esta razón, deben limitarse las expresiones del tipo "todas las anteriores" o "ninguna de las anteriores". La mayoría de los estudiantes inteligentes conocen que las respuestas categóricas de este tipo son casi siempre incorrectas.

Por último, deberían redactarse al menos el doble (40, por ejemplo) de los ítems que constituirán el test final (20), puesto que muchos serán descartados en el proceso de revisión de expertos y el análisis estadístico ulterior.

6.3. Revisión de expertos

La mayoría de los autores recomiendan que los ítems preliminares sean revisados por jueces expertos. Es conveniente que estos jueces tengan experiencia en construcción de pruebas, en el dominio o constructo a medir (autoeficacia, por ejemplo) y en la población a la cual se dirige el test (adolescentes, por ejemplo). Los tres aspectos esenciales que los expertos deben evaluar en cada ítem son:

a. Claridad semántica y corrección gramatical.

b. Adecuación al nivel de comprensión de la población meta.

c. Congruencia con el constructo o dominio medido.

Este último es el principal parámetro y hace referencia al grado de consistencia que debe existir entre un ítem particular y los constructos a medir por el test. Respetar este parámetro

contribuirá significativamente a la confiabilidad y validez de las puntuaciones del test a construir (Oesterlind, 1990).

Los procedimientos empíricos que se utilizan para el juicio de expertos acerca de la calidad de los ítems son los mismos que fueron descritos en el capítulo, de validez (en el apartado de evidencia relacionada con el contenido). Por lo general se emplean escalas numéricas para que los jueces evalúen la calidad y consistencia de los ítems y se descartan aquellos con puntuaciones medias más bajas y con escaso grado de acuerdo, respectivamente. Pueden utilizarse estadísticos de concordancia, tales como el coeficiente kappa mencionado en el capítulo 3 de confiabilidad de los tests. Se recomienda que los ítems seleccionados sean aquellos que, al menos, un 60% de los jueces consideran meritorios (Herrera Rojas, 1998). Es útil también incluir preguntas adicionales sobre los ítems (sobre su facilidad de comprensión, por ejemplo) que faciliten una redacción más adecuada de algunos de ellos.

No deberíamos confiar exclusivamente en el juicio de los expertos y siempre es conveniente realizar una prueba piloto en una muestra pequeña, con el objetivo de corroborar empíricamente que los ítems sean claros y comprensibles para la población meta del test.

6.4. Análisis factorial y de ítems

En tests que miden constructos psicológicos (aptitudes, rasgos de personalidad, intereses, actitudes) el procedimiento esencial y recomendable para construir escalas confiables y con significado teórico es el análisis factorial. También existen otros métodos de análisis de los ítems de un test que se ocupan básicamente de dos aspectos: la distribución de las puntuaciones de cada ítem y la relación estadística entre el ítem y la prueba total (Herrera Rojas, 1998). Estos últimos métodos deberían utilizarse cuando estamos desarrollando una prueba de rendimiento o como procedimiento complementario al análisis factorial. Nos ocuparemos en primer lugar del análisis factorial y, posteriormente, del análisis de ítems.

a) Análisis factorial

El paso decisivo para verificar la unidimensionalidad de cualquier escala, así como para seleccionar y otorgar significado teórico a un conjunto inicial de ítems heterogéneos, es el análisis factorial (Martínez Arias, 1995). Tal como se definió en el capítulo de validez, el análisis factorial es un método para agrupar las variables (ítems, por ejemplo) que se correlacionan fuertemente entre sí, y cuyas correlaciones con las variables de otros agrupamientos (factores) es menor (Aiken, 2003). Según Kline (2000), el análisis factorial es un método estadístico en el cual la variabilidad de las puntuaciones de un conjunto de variables es explicada por un número más reducido de dimensiones o factores.

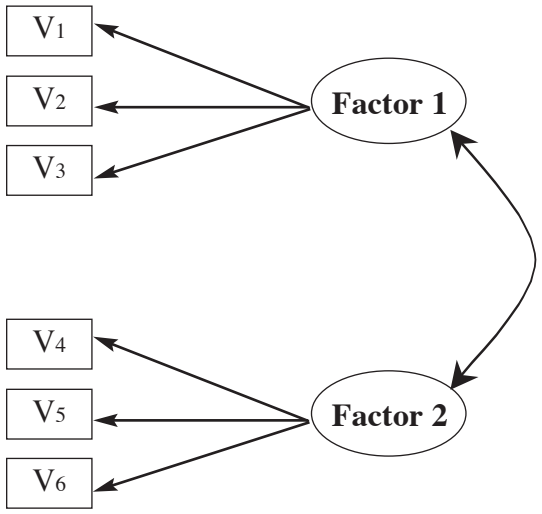
El análisis factorial es, recapitulando, una técnica de reducción de datos cuyo objetivo primordial es agrupar un conjunto de variables en pocas dimensiones que expliquen la mayor cantidad de variabilidad de respuesta. Mediante éste, por ejemplo, un gran número de problemas a resolver (ítems) pueden reducirse a un número pequeño de factores (aptitud verbal, numérica, espacial, por ejemplo) que confieran un significado teórico a la medición. Cada una de estos factores agrupa los problemas (ítems) altamente correlacionados entre sí y que son, al mismo tiempo, relativamente independientes de los restantes factores. Organizaremos la exposición siguiente en relación con las decisiones estratégicas que deben tomarse durante el desarrollo de un análisis factorial.

Análisis factorial exploratorio y confirmatorio

Una distinción inicial importante es la que debe realizarse entre el análisis factorial exploratorio (AFE) y el confirmatorio (AFC). En el primero se extraen e interpretan posibles factores que explican la covariación entre las variables (ítems) sin una estructura teórica previa conjeturada de modo explícito. De esta manera se procedió en el proceso de construcción del Cuestionario de Intereses Profesionales (Fogliatto, 1991), donde se redactó un gran número de ítems relacionados con carreras (sin una teoría explícita) para, posteriormente, identificar los factores (escalas del test) subyacentes y explicativos de la variabilidad de respuesta al test.

En el análisis factorial confirmatorio, por el contrario, la estructura factorial se define a priori sobre la base de un modelo teórico (teoría de los cinco factores de personalidad o teoría de los tres estratos de la inteligencia, por ejemplo). Las hipótesis previas del AFC especifican las relaciones entre los ítems y los factores, como puede apreciarse en la figura siguiente. Posteriormente debe verificarse que esta estructura teórica suministra un buen ajuste a los datos empíricos. Este ajuste se verifica si los valores de los parámetros estimados reproducen tan estrechamente como sea posible la matriz observada de covarianza (Kahn, 2006).

Figura 6.1. Modelo teórico de un hipotético AFC



Como puede observarse en la figura anterior, y basándose en la teoría previa, el investigador ha establecido que los ítems (V) 1, 2 y 3 se asocian con el factor 1 así como los ítems 4, 5 y 6 lo hacen con el factor 2. Asimismo, el modelo teórico establece que

los ítems 1, 2 y 3 no se relacionan con el factor 2 y los ítems 4, 5 y 6 no se relacionan con el factor 1. Se postula también que los factores 1 y 2 están intercorrelacionados, pero sin preestablecer ninguna dirección causal, puesto que la flecha que los une es bidireccional.

Posteriormente el análisis intenta verificar cuán bien se ajustan los resultados observados al modelo teórico preestablecido mediante métodos (Máxima Verosimilitud, por ejemplo) y estadísticos, tales como: chi cuadrado (χ^2), la diferencia en χ^2 entre los modelos comparados, el índice de bondad del ajuste (GFI) y el índice de ajuste comparativo (CFI), entre otros adicionales (RMSEA, por ejemplo).

Los valores de estos estadísticos de bondad del ajuste (CFI, GFI) varían por lo general entre 0 y 1, donde 1 indica un ajuste perfecto. Valores superiores a 0,9 sugieren un ajuste satisfactorio entre las estructuras teóricas y los datos empíricos y valores de 0,95 o superiores, un ajuste óptimo. El chi cuadrado debe ser no significativo para indicar un buen ajuste de los datos observados al modelo teórico de medición. Esto es así porque un valor significativo de χ^2 implica que la estructura del modelo teórico propuesto es significativamente diferente de la indicada por la matriz de covarianza de los datos. No obstante, este último estadístico es sensible al tamaño muestral (cuanto mayor es la muestra mayor será el chi cuadrado) y debe interpretarse con precaución. Cuando se comparan diferentes modelos teóricos, la reducción significativa en chi cuadrado de un modelo con respecto a otro también sugiere un ajuste más adecuado a los datos (Tabachnick y Fidell, 2001).

En un estudio realizado con el WAIS (Taub, 2001) se evaluó el ajuste de tres modelos teóricos alternativos: a) una estructura con un solo factor de inteligencia subyacente al instrumento, b) la estructura clásica de dos factores (verbal y de ejecución), y c) una solución de cuatro factores (comprensión verbal, organización perceptual, memoria de trabajo y velocidad de procesamiento). Los resultados de la investigación demostraron que la estructura de cuatro factores era la que mejor se ajustaba a los datos empíricos y, por consiguiente, no tenía sentido continuar interpretando las puntuaciones del WAIS del modo convencional, utilizando el puntaje de los subtests Verbal y de Ejecución.

Tabla 6.1. Índices de ajuste para tres modelos teóricos del WAIS

Modelo	χ^2	CFI	χ^2 dif.
1. Cuatro factores	81,78	0,986	
2. Dos Factores	147,50	0,943	65,72*
3. Un factor	152,92	0,940	71,14*

Como puede observarse, los índices de la tabla precedente indican que el ajuste del modelo teórico de cuatro factores es el mejor. El CFI es óptimo (>0,95) y el valor de χ^2 es significativamente menor al de los restantes modelos de comparación, el de dos factores (verbal y ejecución) y de uno (factor *g*).

El aprendizaje del análisis confirmatorio debe ser facilitado por cursos específicos y la lectura de textos especializados, como el autorizado volumen de Thompson (2004). En la actualidad existen programas estadísticos, tales como EQS o AMOS, que permiten realizar este análisis sin mayores dificultades.

En la construcción de un nuevo test, los autores comienzan casi siempre por un análisis factorial exploratorio (aun cuando partan de una teoría previa) y, para una validación ulterior, emplean el análisis confirmatorio, que, al igual que el análisis de senderos, sólo se emplea en fases tardías de investigación cuando se posee una teoría bien establecida como basamento (Aron y Aron, 2001). Por consiguiente, los próximos apartados se refieren fundamentalmente al análisis factorial exploratorio.

Tamaño de la muestra

El análisis factorial debe conducirse empleando muestras grandes, de aproximadamente 300 individuos, para obtener resultados útiles y relativamente estables (Tabachnick y Fidell, 2001). Se debería contar idealmente con 10 participantes por variable (ítem en el caso de tests) y como mínimo con 5 por ítem (Nunnally, 1991). Cuando se trata de muestras muy grandes, se recomienda conducir un análisis factorial diferenciado para cada sexo (Kline, 2000). También es importante que la muestra sea lo más heterogénea posible en relación con los constructos

relacionados con el test. De este modo, si alguien desea construir un inventario de intereses vocacionales para estudiantes del ciclo de especialización del secundario, la muestra debería contemplar un número razonable de adolescentes de cada una de las orientaciones de ese nivel educativo, puesto que la utilización de una sola especialidad podría sesgar los resultados (Reise, Waller y Comrey, 2000).

Factibilidad del análisis factorial

Luego de administrar el test a la muestra de investigación, y antes de emprender el análisis factorial, debe determinarse si los ítems están suficientemente interrelacionados para que este método pueda aplicarse provechosamente. Existen algunas pruebas estadísticas que pueden emplearse con esa finalidad. Las más usadas son el test de esfericidad de Bartlett y la medida de adecuación muestral de Kaiser-Mayer-Olikin (KMO). El KMO se interpreta de manera semejante a los coeficientes de confiabilidad, vale decir, con un rango de 0 a 1 y considerando como adecuado un valor igual o superior a 0,70, el cual sugiere una interrelación satisfactoria entre los ítems. Si éste es el caso para los datos que se poseen es lícito utilizar el análisis factorial en sus diferentes variantes.

Del mismo modo, como casi todos los métodos multivariados, el análisis factorial posee supuestos estadísticos exigentes que deben respetarse y cuya violación puede conducir a resultados equívocos. Entre otros supuestos (tales como el tamaño de la muestra y la interrelación de las variables a analizar), antes de emprender un análisis factorial debe verificarse la distribución normal de las puntuaciones en las variables incluidas en el análisis (normalidad) y la existencia de casos con puntuaciones marginales en esas variables (*outliers*). En Tabachnick y Fidell (2001) se presenta un tratamiento muy exhaustivo de los supuestos de los diferentes métodos multivariados, incluyendo el análisis factorial.

Métodos de extracción de factores

El paso siguiente es decidir el método a emplear para la extracción de factores. Si bien existen varios disponibles en los programas usuales de análisis estadístico (SPSS, por ejemplo), en la práctica los métodos más usados en el análisis factorial exploratorio son dos: componentes principales y ejes (o factores) principales.

Es necesario realizar algunas acotaciones previas que nos permitan comprender las diferencias esenciales entre ambos métodos. Tal como afirmamos precedentemente, el análisis factorial es un método analítico de condensación de la varianza total de respuesta a las variables (ítems en el caso de un test psicológico). Esta varianza tiene tres elementos principales: a) la varianza común (o comunalidad), que es la proporción de varianza de las variables que es explicada por los factores comunes; b) la varianza específica, que es el porcentaje de varianza particular de cada variable; y, c) la varianza de error, que es el porcentaje de varianza no explicada, atribuible al error de medición.

El método de componentes principales explica la mayor cantidad de varianza posible en los datos observados. Por consiguiente, este método analiza la varianza total asociada a las variables, incluyendo la varianza específica y la varianza de error (Juan-Espinosa, 1997). El método de ejes principales, en cambio, sólo analiza la varianza que las variables tienen en común o covarianza, excluyendo la específica y la atribuible al error de medida (Tabachnick y Fidell, 2001).

El método de componentes principales es más fácil de interpretar que el de ejes principales y tal vez en eso radique su mayor popularidad, particularmente cuando se analiza un conjunto grande de ítems para desarrollar nuevas escalas o inventarios (Merenda, 1997). Si el análisis factorial se emplea para obtener una solución teórica no contaminada por la varianza de error y específica, el método de ejes principales es la alternativa más adecuada (Tabachnick y Fidell, 2001; Costello y Osborne, 2005).

No obstante, cuando los tests poseen confiabilidades adecuadas, las diferencias entre las soluciones factoriales obtenidas por cada método suelen ser poco importantes (Kline, 2000).

Número de factores a extraer

La extracción del número correcto de factores es una de las decisiones más problemáticas del análisis factorial (Tabachnick

y Fidell, 2001). El empleo de un único criterio puede llevar a sobrestimar o subestimar el número real de factores y, por ese motivo, se recomienda emplear un conjunto de criterios para identificar el número de factores subyacentes en las escalas psicológicas. Si la alternativa es extraer más factores o menos (sobre y sub-extracción), la sobre-extracción es menos riesgosa puesto que conlleva menos error en la medición (Reise, Waller y Comrey, 2000). No obstante, la decisión acerca del número de factores a extraer debería sustentarse siempre en evidencia empírica.

Un método muy empleado, y que aparece por defecto en el programa SPSS, es la regla Kaiser de extracción de factores con autovalores (*eigenvalues*) superiores a 1. El cuadrado de una correlación entre una variable y un factor es la proporción de varianza explicada por esa variable. Si se suman todos los cuadrados de los pesos factoriales de las variables en un factor (columna de la matriz factorial) obtenemos el autovalor de ese factor, que expresa la magnitud de varianza explicada por ese factor. El punto de corte de 1 se fija porque las variables están estandarizadas con la varianza igual a 1 y sería inadecuado interpretar un factor que explique menos varianza que la explicada por una variable particular (Kahn, 2006). Si dividimos el autovalor de un factor por el número de variables y multiplicamos ese valor por 100 obtenemos el porcentaje de varianza explicada por ese factor particular. El inconveniente principal de esta regla es que generalmente conduce a la extracción de demasiados factores, particularmente en tests con muchos ítems (50 ó más).

Otro criterio de extracción es el porcentaje de varianza explicada por la estructura factorial obtenida (varianza acumulada del número de factores extraídos). En este caso se recomienda que la solución factorial explique, al menos, un 50% de la variabilidad total de respuesta al test (Merenda, 1997). El porcentaje de explicación de varianza puede ser una condición necesaria, pero en la práctica es un criterio poco decisivo, puesto que podemos tener varias soluciones factoriales alternativas con porcentajes adecuados de varianza explicada y, por consiguiente, no sabremos por cuál optar. En todo caso, la regla Kaiser y el porcentaje de varianza explicada son procedimientos complementarios aunque no esenciales en la mayoría de los casos.

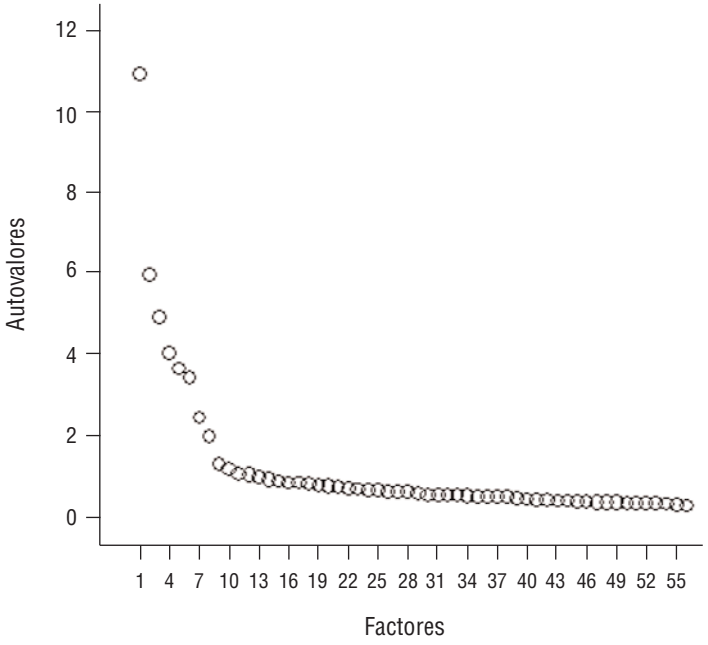
El criterio de extracción de factores más empleado en la actualidad es el denominado *scree test* o *scree plot* (gráfica *scree*) (Cattell, 1966). El *scree test* es una representación gráfica del tamaño de los autovalores y ayuda a identificar el número óptimo de factores que se deberían extraer. En el eje vertical u ordenada se representan los autovalores, y en el horizontal o abscisa, el número de factores. Sobre la gráfica resultante se traza una línea recta base a la altura de los últimos autovalores (los más pequeños) y aquellos que queden por encima de esa línea base indicarán el número de factores a retener. Cattell (1966) denominó a esta gráfica "*scree*" por su parecido al perfil de la falda de una montaña, donde los residuos rocosos de la base son comparables a los factores irrelevantes de la solución, metafóricamente no sólidos.

El *scree* test es un procedimiento con un componente de subjetividad pero se ha verificado la adecuada confiabilidad del mismo (Kline, 2000). En general, el punto de corte para el número de factores a extraer está determinado por el primer cambio de pendiente en la gráfica (Juan-Espinosa, 1997). Los autovalores residuales se ubican a la derecha del gráfico, formando una planicie de poca inclinación. En cambio, los autovalores que explican la mayor parte de la varianza se ubican en la parte izquierda formando una fuerte pendiente. Se recomienda inspeccionar el gráfico de izquierda a derecha hasta localizar el punto de inflexión en que los autovalores dejan de formar una pendiente y comienzan a generar una caída de poca inclinación.

En la figura 6.2. puede apreciarse que si bien el análisis factorial realizado en el Inventario de Autoeficacia para Inteligencias Múltiples (IAMI) indicaba la existencia de 12 factores utilizando la regla Kaiser, el *scree test* sugiere que solamente 8 deberían ser interpretados puesto que la caída o declive de la gráfica se interrumpe a partir del noveno autovalor. En efecto, entre el noveno y décimo autovalor no existe declive y se comienza a configurar una línea plana.

El *scree test* es confiable en la mayoría de los casos, pero algunas veces resulta dificultoso determinar el número exacto de factores con un mero examen visual del gráfico, en especial

Figura 6.2. Scree test del análisis factorial del IAMI



cuando los factores son numerosos o existen varios cambios de dirección en la pendiente.

Horn (1965) propuso otro método, el análisis paralelo, que parece ser una de las mejores alternativas para decidir el número de factores a extraer. Este análisis genera autovalores de una matriz de datos aleatorios, pero con el mismo número de variables y casos que la matriz original. Si bien el análisis paralelo no puede ejecutarse desde los programas estadísticos usuales (SPSS, SAS, por ejemplo), Thompson y Daniel (1996) desarrollaron una sintaxis que puede correrse de manera sencilla desde SPSS.

En el análisis paralelo se compara en una tabla el autovalor de cada factor en los datos reales con el autovalor correspondiente de los datos aleatorios. Para decidir el número de factores a extraer se identifica el autovalor de los datos reales con magnitud superior al autovalor de los datos simulados. La lógica del procedimiento es que los factores reales que explican más varianza que los aleatorios deben interpretarse (Kahn, 2006).

Tabla 6.2. Autovalores para datos reales y ordenados al azar

Factor	Autovalores datos reales	Autovalores datos aleatorios
1	2,776	1,746
2	0,838	1,052
3	0,376	0,790
4	0,121	0,365

Como puede observarse en la tabla precedente, el autovalor 2,776 es más elevado que el autovalor 1,746 y, por consiguiente, ese factor debería ser extraído. Los restantes factores presentan autovalores superiores en la columna de datos ordenados al azar y no deben ser interpretados. Sintetizando, y de acuerdo al análisis paralelo, en el ejemplo sólo existe un factor interpretable. El Dr. Ruben Ledesma, de la Universidad Nacional de Mar del Plata, ha desarrollado y difundido un *software* estadístico de libre acceso (VISTA). Entre otras prestaciones, este programa permite correr el análisis paralelo, incluyendo atractivos gráficos y un entorno amigable para el usuario (Ledesma y Valero-Mora, 2007).

Rotación de factores

El resultado inicial del análisis factorial (antes de la rotación) es una matriz factorial, es decir, la matriz de correlaciones de las variables con los factores. Cuando se trata de tests con ítems dicotómicos (verdadero-falso; correcto-incorrecto) debe analizarse de la matriz de correlaciones tetracóricas. En efecto, el coeficiente de correlación tetracórica es apropiado para las variables dicotomizadas (Martínez Arias, 1995). En algunos programas estadísticos puede obtenerse directamente la matriz de correlación tetracórica. El paquete SPSS posee macros (sintaxis especial) que permite estimar dicha matriz. Esta matriz factorial

inicial es difícil de interpretar y, en casi todos los casos en los que se extrae más de un factor, es indispensable obtener una matriz adicional de factores rotados.

Por consiguiente, luego de extraer los factores iniciales, éstos son sometidos a un procedimiento denominado rotación (cuando hay más de un factor en la solución). El término rotación proviene de la representación gráfica y geométrica del análisis factorial; en efecto, los factores pueden representarse como ejes de referencia y los pesos factoriales (correlaciones) de cada variable indicarse en los ejes correspondientes (Kerlinger y Lee, 2002). La rotación intenta que la solución factorial se aproxime a lo que se denomina estructura simple, esto es, que cada ítem tenga una correlación lo más próxima a 1 que sea posible con uno de los factores, y correlaciones próximas a 0 con los restantes factores. El investigador rota los factores con la finalidad de eliminar las correlaciones negativas importantes y reducir el número de correlaciones de cada ítem en los diversos factores (Anastasi y Urbina, 1998).

Naturalmente, nunca encontramos en los datos empíricos estructuras simples sino una solución aproximada a ese concepto teórico. Las rotaciones pueden ser ortogonales u oblicuas y dos métodos muy empleados son Varimax y Promax, respectivamente, aunque hay otros disponibles en los programas estadísticos (SPSS, por ejemplo).

Las soluciones provistas por los métodos de rotación oblicua son más congruentes con la estructura de las variables psicológicas que, en general, se encuentran intercorrelacionadas. La ortogonalidad absoluta es sólo teórica y, a los fines prácticos, se interpreta que una solución es ortogonal cuando las correlaciones entre los factores son inferiores a 0,32. Tabachnick y Fidell (2001) proponen realizar una rotación oblicua inicial como filtro (Promax, por ejemplo), y obtener la matriz de correlación entre los factores. Si observamos alguna correlación superior a 0,32 entre los factores deberíamos escoger una rotación oblicua, y en caso contrario, una ortogonal.

Las rotaciones colocan a las variables más cerca de los factores diseñados para explicarlas, concentran la varianza de las variables en menos factores y, en general, proporcionan un medio para facilitar la interpretación de la solución factorial obtenida. En la actualidad existen varios algoritmos ejecutables en los paquetes estadísticos que generan la matriz rotada sin recurrir a procedimientos gráficos de rotación (Thompson, 2004). Con estas operaciones algebraicas (multiplicación de los coeficientes no rotados por un conjunto de constantes derivadas mediante funciones trigonométricas) la estructura de la matriz factorial se modifica y es más sencilla de interpretar, debido al incremento de las correlaciones positivas extremas (bajas y altas), aproximándose a la estructura simple ideal que mencionáramos. Los procedimientos analíticos de rotación han reemplazado a los geométricos por su sencillez (las ejecutan los programas informáticos) y objetividad (es más difícil alcanzar resultados idénticos entre varios investigadores cuando se rota gráficamente).

En la tabla 6.3 se presenta la matriz factorial de diez subtests, sin rotar, y en la tabla 6.4 los factores rotados; el examen de las mismas facilitará la comprensión de los conceptos desarrollados.

Tabla 6.3. Matriz factorial sin rotar de las escalas de un test de inteligencia

Subtests	Factor I	Factor II
Vocabulario	0,74	0,54
Analogías	0,62	0,37
Frases Incompletas	0,66	0,41
Frases en desorden	0,30	0,21
Comprensión	0,68	0,48
Suma	0,20	-0,51
Multiplicación	0,43	-0,53
Cálculos aritméticos	0,51	-0,46
Ecuaciones	0,44	-0,35
Completamiento de series numéricas	0,31	-0,23

Tabla 6.4. Matriz factorial rotada de los mismos datos

Subtests	Factor I	Factor II
Vocabulario	0,90	-0,05
Analogías	0,73	0,01
Frases Incompletas	0,81	0,00
Frases en desorden	0,36	-0,02
Comprensión	0,85	-0,03
Suma	-0,08	0,52
Multiplicación	0,06	0,63
Cálculos aritméticos	0,17	0,66
Ecuaciones	0,15	0,52
Completamiento de series numéricas	0,12	0,36

Como puede apreciarse en el ejemplo, es más sencillo interpretar la matriz rotada que la matriz factorial sin rotar, en la que es más difícil determinar cuáles variables se relacionan con cada factor debido a que las correlaciones de los subtests con ambos factores son relativamente semejantes. En la matriz rotada puede identificarse claramente qué cinco subtests correlacionan fuertemente con un factor que podríamos denominar "Verbal", mientras que los cinco restantes configuran otro factor que se puede interpretar como "Numérico".

Para obtener una solución aproximada a la estructura simple, las correlaciones entre un ítem y un factor deberían ser de 0,40, al menos, y no debería existir una correlación superior a 0,30 de esa variable con otro factor. De no ser así, estaríamos reteniendo ítems complejos, así como soluciones factoriales insatisfactorias y difíciles de interpretar.

Hay que considerar que si hemos empleado una rotación oblicua (promax, por ejemplo) no obtendremos sólo una matriz rotada (como en la rotación ortogonal) sino dos, que se denominan matriz de estructura y de configuración. En la matriz de estructura se presentan las correlaciones de cada variable con el factor o coeficientes estructurales. En cambio, en la matriz de configuración los coeficientes observados son análogos a los coeficientes beta del análisis de regresión múltiple. Los coeficientes de configuración indican la importancia relativa de cada factor para explicar el puntaje individual en cada variable, controlando los restantes factores. La mayoría de los investigadores analizan la matriz de configuración debido a su mayor facilidad de interpretación (Tabachnick y Fidell, 2001), pero se recomienda atender a ambas matrices para una interpretación más adecuada de los resultados (Thompson, 2004).

Interpretación de los factores

La tarea final del análisis factorial es interpretar y nominar los factores. Esto se logra examinando el patrón de correlaciones bajas y altas de cada variable con los distintos factores y, en especial, utilizando el conocimiento teórico que se posea acerca de las variables incluidas en el análisis. Como vimos en el ejemplo anterior (análisis factorial a nivel de subtests o de segundo orden), los subtests de Analogías, Frases Incompletas, Vocabulario y Comprensión poseen correlaciones superiores a 0,40 con uno de los factores. Puesto que estos subtests son todos de tipo verbal, este factor podría denominarse "Aptitud Verbal". Del mismo modo, Suma, Multiplicación y Ecuaciones se correlacionan fuertemente con el otro factor, que podríamos denominar "Aptitud Numérica". Se recomienda que cada factor posea, al menos, cuatro ítems con correlaciones iguales o superiores a 0,40 para ser interpretado (Glutting, 2002).

Algunas consideraciones finales

Cuando los factores obtenidos están intercorrelacionados es posible continuar el análisis factorial y obtener "factores de factores" (Juan-Espinosa, 1997), tal como vimos en el ejemplo anterior. En otros términos, podemos realizar un análisis factorial y descubrir factores oblicuos de primer orden para después analizar factorialmente la matriz de correlación entre los factores y derivar factores de segundo orden o superior (Anastasi y Urbina, 1998). Éste es el caso del 16PF-5 (Russell y Karol, 2000), por ejemplo, donde los factores (rasgos) iniciales son 16, pero un nuevo análisis redujo el modelo a 5 factores de segundo orden semejantes a los empleados por el Inventario NEO-PI-R (Costa

y Mc Crae, 1999), que son estabilidad, extroversión, apertura, amabilidad y responsabilidad. Algunos programas de computación realizan este análisis en forma sencilla y en un solo paso (Kerlinger y Lee, 2002).

En general, el análisis factorial exploratorio debe complementarse con una estrategia confirmatoria posterior (Aron y Aron, 2003). El empleo exclusivo del análisis factorial exploratorio puede conducirnos a obtener estructuras meramente empíricas, dependientes de las muestras e ítems seleccionados, y no replicables con facilidad. Como afirma Eysenck (1981), "*el análisis factorial es un buen servidor pero un mal amo*", Este investigador partió de teorías biológicas de la personalidad (y la inteligencia) y, posteriormente, utilizó el análisis factorial para buscar una confirmación de las hipótesis derivadas de esos modelos teóricos.

b) Análisis de ítems

Además del análisis factorial, existen otros procedimientos complementarios que son útiles para analizar la calidad de los ítems de un test. Para implementar estos métodos de análisis de ítems se recomienda administrar el test a una muestra por lo menos cinco veces superior a la del número inicial de ítems y, como se mencionó anteriormente, es deseable contar aproximadamente con el doble de ítems de los que aparecerán en la versión definitiva del test.

La determinación del número muestral necesario para realizar análisis de ítems y, en general, todos los estudios psicométricos esenciales (confiabilidad, validez) de un test, es un motivo de fuerte discusión entre los psicometristas, en particular en ciertas áreas, como la neuropsicología, donde obtener muestras extensas es altamente problemático porque los tests empleados son de administración individual o los individuos son difíciles de conseguir (Charter, 1999).

Una muestra de 300 personas para análisis de ítems es un estándar deseable pero no garantiza obtener propiedades psicométricas adecuadas de la escala (coeficiente alfa de 0,80 o superior, por ejemplo). Otros factores intervinientes, tales como el

entrenamiento de los examinadores, el uso de formatos de prueba que aseguren mayor variabilidad de las respuestas o la heterogeneidad de la muestra pueden incrementar los valores de confiabilidad y validez y compensar tamaños muestrales inferiores (Pajares, Hartley y Valiante, 2001).

El procedimiento más empleado en el análisis de ítems es la correlación de cada ítem con el puntaje total de la prueba. Este índice permite identificar la capacidad del ítem para discriminar (diferenciar) entre los individuos que poseen "más" un rasgo y los que poseen "menos" de ese rasgo. Como afirman Anastasi y Urbina (1998), la capacidad discriminativa de un ítem se refiere al grado en que éste diferencia adecuadamente entre los examinados en relación con el rasgo o dominio que el test pretende medir. Si el test posee varias escalas o subtests, cada ítem debe correlacionarse también con el puntaje total de esa sección.

El estadístico usual es el coeficiente producto momento de Pearson (r) o punto-biserial si se trata de ítems dicotómicos (correcto/incorrecto, por ejemplo). El coeficiente de correlación punto-biserial se utiliza cuando una de las variables es dicotómica (sólo puede tomar dos valores) y la otra continua (puntaje total en una escala, por ejemplo) (Velandrino, 1998).

Los ítems con correlaciones no significativas o bajas con el puntaje total (inferiores a 0,30) deben eliminarse o revisarse. Los reactivos que se retendrán después de este análisis inicial serán, seguramente, los menos ambiguos, ni excesivamente fáciles ni dificultosos y más relacionados con el constructo o dominio medido por el test (Nunnally y Bernstein, 1995). En las escalas *likert*, aquellos ítems que la mayoría de los individuos responde utilizando sólo las alternativas extremas (en total desacuerdo o en total acuerdo, por ejemplo) están indicando que carecen de capacidad discriminativa (Bandura, 2001).

En pruebas educativas de opción múltiple es recomendable también obtener las correlaciones de cada una de las alternativas de respuesta (distractores) con el puntaje total. Estos resultados pueden orientarnos en las decisiones sobre los ítems, puesto que permiten identificar aquellas alternativas que, sin ser la opción correcta, son elegidas por los examinados con alto nivel de maestría del dominio. Los mejores distractores serán

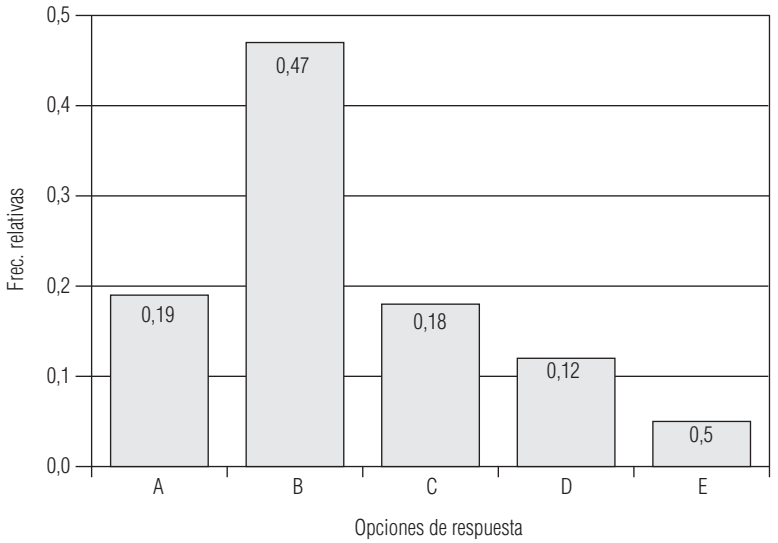
aquellos que obtengan correlaciones negativas con los puntajes de la prueba, es decir que sean seleccionados por quienes tienen puntajes bajos en la prueba (Herrera Rojas, 1998).

En las pruebas de habilidades o rendimiento (con ítems dicotómicos del tipo correcto/incorrecto) otro indicador importante es el índice de dificultad de cada ítem, vale decir, el porcentaje de personas que responden acertadamente al ítem analizado. El índice de dificultad de los ítems tiene un rango de 0 a 1 y se simboliza como p. Un ítem cuyo p es 0 está indicando que ningún individuo lo contestó correctamente y un ítem con p igual a 1 es aquel que todos los individuos respondieron de manera correcta. El valor óptimo de p para un ítem depende de varios factores, tales como los objetivos del test y la cantidad de opciones de respuesta. Si el propósito del test es identificar sólo un porcentaje reducido de los mejores postulantes a un curso o empleo, por ejemplo, entonces los ítems de prueba deberían ser los suficientemente difíciles y tener un p bajo (0,20, por ejemplo). Cuando la selección de personas no es el propósito esencial, son deseables ítems con p entre 0,30 y 0,70, vale decir ni excesivamente difíciles ni fáciles (Aiken, 2003; Kaplan y Saccuzzo, 2006).

La proporción de acierto es un estimador adecuado de la dificultad de un ítem. Sin embargo, esta información debe ser complementada con la distribución de frecuencias para todas las opciones de respuesta (cuando se trata de pruebas de elección múltiple) y las estimaciones de proporción para diferentes rangos de puntuación en la prueba total. La figura siguiente presenta la distribución de frecuencias de las opciones de respuesta a un ítem, en una prueba de conocimientos matemáticos compuesta de 100 preguntas con cinco opciones de respuesta, aplicada a una muestra de 981 estudiantes.

Puede observarse que el ítem en cuestión resulta de dificultad media para esa muestra y fue contestado correctamente por el 47% de los estudiantes, puesto que la opción B es la correcta. También se infiere de la figura anterior que los distractores no fueron elegidos en la misma proporción. El distractor E, por ejemplo, sólo fue elegido por el 5% de los examinados. La distribución de frecuencias de las opciones de respuesta es útil para identificar los distractores que son elegidos por una alta proporción de personas, así como aquellos fácilmente descartables incluso para las personas que poseen baja maestría del dominio (Kehoe, 1995).

Figura 6.3. Distribución de frecuencias relativas por alternativa de respuesta en una prueba de conocimientos



En el caso de la calificación de pruebas educativas de opción múltiple, debido a que los examinados pueden emitir algunas respuestas correctas por adivinación, existen diversos procedimientos que penalizan esta situación (Kaplan y Saccuzzo, 2006). Una de las fórmulas para obtener esta puntuación corregida es:

$$R - \frac{W}{k-1}$$

Los términos de la fórmula precedente son R (número de respuestas correctas), W (número de respuestas equivocadas) y k (número de opciones de respuesta de cada ítem). Las respuestas omitidas (preguntas no respondidas) no se incluyen puesto que no proporcionan crédito ni penalización.

Supongamos que un estudiante responde a una prueba de 10 ítems con tres opciones de respuesta y posee cuatro respuestas correctas, cuatro erróneas y deja dos ítems sin responder. Aplicando la fórmula de corrección de las puntuaciones tendríamos:

$$1 \pm \frac{4}{3 \pm 1} =$$

Es decir que la puntuación real (corregida por azar) en este caso sería de 2 para el estudiante en cuestión.

Existe una fórmula alternativa que atenúa el efecto del descuento aplicado a las respuestas emitidas por azar (Martínez Arias, 1995):

$$R - \frac{W}{2\ (k \pm 1)}$$

Donde, reemplazando los términos ya explicados por los valores del ejemplo anterior, tendríamos:

$$4 - \frac{4}{2\ (3 - 1)} =$$

Debe destacarse que, en el caso de las pruebas referidas a criterio, se analizan y seleccionan los ítems de manera diferente. El análisis se realiza comparando los resultados de un grupo antes de aplicar un programa educativo y después del mismo, o comparando dos grupos similares, uno de ellos que recibió capacitación y el otro no. Al calcular el índice de dificultad, los resultados esperados son ítems con alta dificultad para los grupos que no han pasado por el proceso de aprendizaje, y baja dificultad para los que han sido sometidos al programa de instrucción. En cuanto al índice de discriminación, obtenido por la comparación entre grupos, se espera máxima discriminación entre los grupos, y mínima entre los individuos de un mismo grupo (Martínez Arias, 1995).

Una vez obtenidos los resultados estadísticos sobre el comportamiento de cada ítem se podrán tomar decisiones acerca de cuáles de ellos deben integrar la forma final de la prueba y hacer estimaciones de confiabilidad y validez mediante algunos de los procedimientos presentados en los capítulos precedentes.

Un objetivo esencial del análisis de ítems es obtener pruebas homogéneas (internamente consistentes), en las que todos los reactivos se relacionen con un núcleo común de medición que es el constructo o dominio, información que se obtiene mediante el coeficiente alfa de Cronbach u otro indicador de consistencia interna. El conjunto de ítems seleccionado es analizado utilizando alfa o algún coeficiente semejante, y debemos asegurarnos valores de 0,80 o superiores. Los programas computarizados permiten identificar fácilmente aquellos ítems con correlaciones bajas con el puntaje total que pueden eliminarse para incrementar el valor de alfa. En el apéndice se ilustra este último procedimiento de análisis.

Si bien un coeficiente alfa adecuado es una condición necesaria de unidimensionalidad, esta propiedad debe ser determinada principalmente por el análisis factorial (Goldberg, 1999). En efecto, es común observar escalas con coeficientes alfa de 0,80 o más que, no obstante, poseen una estructura multifactorial (Kline, 2000). Por consiguiente, la estrategia de análisis de ítems no es recomendable como procedimiento esencial de construcción de un test que mida rasgos latentes, sino para aportar información complementaria acerca de la calidad de los ítems (índices de discriminación y dificultad del ítem).

Esta última aseveración debe entenderse como referida a los tests construidos partiendo de las premisas de la TCT. Por el contrario, cuando los tests son construidos utilizando los modelos TRI, los procedimientos de análisis de ítems son esenciales. El impacto de la TRI es limitado en el mundo hispano (con la excepción de algunos tests educativos) aunque se percibe un crecimiento sostenido de su empleo. Tal como sugiere Goldberg (1999), los criterios de construcción de la teoría clásica se pueden complementar con algunos de los procedimientos del modelo TRI, que revisaremos en el último capítulo. En la Argentina, pocos investigadores trabajan en la construcción de tests utilizando métodos de la teoría de respuesta al ítem. Una notable

excepción es el Test Baires de Aptitud Verbal (Cortada de Kohan, 1998) que comentamos en el capítulo sobre clasificación de tests, así como las investigaciones conducidas por un número reducido de especialistas de la cátedra de Psicoestadística, de la Universidad de Buenos Aires.

El último paso en la construcción de un test es la realización de los estudios psicométricos esenciales que hemos revisado en los capítulos de la sección de normas técnicas (estandarización, confiabilidad, validez). El marco teórico del test y los estudios psicométricos aparecen en el manual del instrumento, al cual se adjuntan los materiales correspondientes (cuadernillos y hojas de respuesta, instrucciones de administración, etc.).

Una alternativa a la construcción de tests es la de adaptar instrumentos elaborados en otros países (casi siempre en los Estados Unidos), problemática que revisaremos en el capítulo siguiente. La decisión de construir un test nuevo o adaptar uno existente es siempre difícil y debería pensarse cuidadosamente, analizando las ventajas y desventajas de cada curso de acción.

7
ADAPTACIÓN DE TESTS A OTRAS CULTURAS

Alberto Fernández

7.1. Por qué adaptar tests

La utilización de tests psicológicos creados en otros contextos culturales es una práctica habitual en todo el mundo. Este fenómeno es particularmente frecuente en los países de las regiones con menor desarrollo científico, tales como Latinoamérica y África. Sin embargo, también ocurre en las naciones más desarrolladas, donde usualmente se utilizan los tests construidos en los Estados Unidos, sin duda el país en el que se producen la mayoría de los instrumentos de medición psicológica.

El uso de un test en un contexto cultural diferente al original genera diversas dificultades. El idioma, la familiaridad con los estímulos del test (ítems) y las diferentes características de las muestras de estandarización son ejemplos de fuentes de posibles sesgos en la medición transcultural de constructos psicológicos.

Existe sesgo en la medición cuando las diferencias individuales en las puntuaciones de un test no reflejan las diferencias reales en un rasgo o habilidad. Tomemos como ejemplo una hipotética investigación en la cual se midiese la capacidad de denominar objetos o animales. Si entre los ítems estuvieran incluidas las figuras de un canguro y un oso koala es más probable que una muestra de estudiantes australianos (lugar en el que viven estos animales) obtenga puntuaciones superiores a las de una muestra de estudiantes nigerianos e incluso a la de muchos de los lectores de este texto. Estos datos no estarían demostrando una mayor capacidad de denominación de los estudiantes

australianos, sino que sugerirían que el indicador empleado en la medición está sesgado. Es decir, existe un elemento diferente (en este ejemplo, la familiaridad con el estímulo presentado) de la capacidad (de denominación en este caso) que influye en el desempeño en la prueba. Por ende, no se estaría midiendo la habilidad en forma equivalente en ambos casos, a pesar de estar utilizando la misma prueba. La equivalencia puede definirse por lo tanto de manera opuesta al sesgo, es decir, se manifiesta cuando las puntuaciones de un test reflejan diferencias entre las personas evaluadas que existen verdaderamente en el rasgo en cuestión.

La existencia de sesgo en un test puede conducir a obtener resultados gravemente erróneos. En una prueba utilizada en psicología clínica, por ejemplo, se podría inferir la presencia de un trastorno de personalidad cuando el rasgo o comportamiento así diagnosticado es normal en la cultura del individuo examinado.

Además de la obvia necesidad de contar con instrumentos adecuados para la práctica psicológica, la adaptación de tests a diferentes culturas obedece a otras razones, de índole científicas y prácticas.

Entre las primeras, es importante tener en cuenta que la mayoría de las teorías psicológicas contemporáneas han sido desarrolladas en el marco de la cultura occidental, más precisamente en universidades norteamericanas. Asimismo, el proceso de validación de dichas teorías se realiza preferentemente en investigaciones que utilizan muestras de jóvenes universitarios de raza blanca.

En la actualidad se reconoce la necesidad de demostrar la "universalidad" de las teorías, si es que ello fuese posible. De este modo, para poder determinar si determinado constructo psicológico existe en otras culturas es necesario contar con tests equivalentes, es decir, que midan el mismo constructo en las diferentes culturas donde va a ser utilizado.

Las razones prácticas se relacionan con la dinámica de la globalización y los fenómenos migratorios, principalmente dentro de los países más desarrollados. Las personas que son evaluadas en un proceso de selección de personal, por ejemplo, provienen de diferentes partes del mundo y, para que esa evaluación

sea justa, es necesario contar con instrumentos adecuados (equivalentes o sin sesgo). Lo mismo ocurre en los casos en los que se administran tests a individuos pertenecientes a minorías étnicas de un país.

7.2. Fuentes de sesgo

Van de Vijver y Tanzer (1997) identificaron diferentes fuentes de sesgo que se describen a continuación.

Sesgo de constructo

Este sesgo se presenta "cuando el constructo medido no es idéntico en diferentes grupos culturales" (Van de Vijver y Tanzer, 1997, p. 264.). Comportamientos morales que en algunas sociedades pueden ser normales en otras pueden constituir un verdadero rasgo de rigidez y asemejarse a una conducta obsesivo-compulsiva. McCrae, Yik, Trapnell, Bond y Paulhus (1998) encontraron, por ejemplo, importantes diferencias entre los perfiles de personalidad de estudiantes canadienses y estudiantes chinos utilizando versiones equivalentes del NEO PI-R. Los estudiantes chinos obtuvieron puntajes significativamente menores en algunas facetas de la escala Extraversión de este último test.

Sesgo metodológico

Este tipo de sesgo reconoce tres formas:

a) El *sesgo de muestra*, que ocurre cuando las muestras son incomparables entre sí. La cantidad de años de escolaridad que poseen los individuos de una muestra constituye una variable determinante en el desempeño de los mismos en un test determinado, especialmente si se trata de un test de ejecución máxima (Heaton, Grant y Matthews, 1991). Los tests de razonamiento lógico o matemático, por ejemplo, presentan una dificultad considerablemente ma-

yor para las personas con baja escolaridad. El nivel sociocultural, la motivación, el sexo y la edad de los sujetos son otras de las variables que pueden hacer incomparables a dos muestras.

Fernández y Marcopulos (2004) compararon los estudios normativos de un test de atención en diez países y advirtieron que una de las principales dificultades para comparar los puntajes residía en las diferencias entre las muestras. Así, observaron que la edad media de los ancianos de la muestra neozelandesa estaba 1,6 desviaciones estándar por encima de la muestra danesa. Sin embargo, en la muestra danesa el 80% de los individuos tenía siete o menos años de educación mientras que la muestra neozelandesa tenía un promedio de diez años de educación. De este modo, no era posible comparar los puntajes de ambas muestras puesto que el nivel educativo (variable que tiene una gran influencia en el desempeño en este test) era muy diferente. Por consiguiente, las diferencias observadas en los puntajes del test posiblemente no reflejaban diferencias reales en la habilidad atencional sino el nivel educativo de los grupos. Es probable que la inclusión de una muestra de individuos daneses con el mismo nivel educativo que los neozelandeses hubiera reducido notablemente esas diferencias.

b) El sesgo en el instrumento puede provenir de las características del test. La familiaridad que los sujetos tengan con los ítems presentados es un aspecto de gran importancia. Algunos estímulos tales como objetos, dibujos, figuras u otros elementos utilizados en algunas culturas no existen en otras o son irrelevantes. El ítem de ejemplo en el subtest de Ordenamiento del WISC-III (Wechsler, 1994) que muestra a una mujer frente a una máquina expendedora de latas de gaseosa tiene muy poco valor en culturas árabes, por ejemplo, o aun en zonas rurales de nuestro país. Las máquinas expendedoras de gaseosas no son comunes en todo el mundo y mucho menos su utilización, lo que puede hacer incomprensible dicha lámina para muchos individuos.

El idioma es otra fuente de sesgo del instrumento. La traducción es un problema considerable y requiere una metodología específica que se explicará más adelante. Los problemas pueden ser aún mayores cuando, por ejemplo, los idiomas son tan distintos como el inglés y el árabe, en los cuales la lectura se realiza de izquierda a derecha y de derecha a izquierda, respectivamente. En algunos tests que incluyen completamiento de frases, comprensión lectora o cancelación de letras, la disposición del texto tiene mucha relevancia. El problema es aún mayor cuando, por ejemplo, el estímulo del test es un caracter del alfabeto, como en el caso del Test del Trazo (TT, Trail Making Test), donde se deben conectar en forma alternada letras y números siguiendo el orden alfabético y un orden numérico ascendente. Incluso entre las lenguas occidentales existen diferencias en los caracteres del alfabeto: el inglés no contiene la “ñ”, el alfabeto sueco contiene las siguientes vocales, a, e, i, o, u, y, å, ä, ö; y en portugués existen la â y la ã, entre otros ejemplos. En el caso del TT, Axelrod, Aharon-Peretz, Tomer y Fisher (2000) han adaptado una versión hebrea de este test utilizando los caracteres propios de ese idioma.

Los materiales o estímulos de respuesta constituyen otra fuente posible de sesgo del instrumento. Las láminas de respuestas del Test de Matrices Progresivas, de Raven (1993), que implican completar una secuencia lógica seleccionando una figura entre un grupo de alternativas, incluyen la figura faltante al final de la segunda fila con lo que se asume una lectura de izquierda a derecha. Este hecho fue observado por Carpenter, Just y Shell (1990), y constituye una severa desventaja para los individuos árabes, quienes involuntariamente van a intentar resolver la prueba de derecha a izquierda, el modo en que se lee su idioma.

c) Finalmente, la última variedad de sesgo metodológico es el de administración. Esta categoría incluye los problemas de comunicación, es decir, dificultades para que el entrevistado entienda las instrucciones del entrevistador ya sea por el tipo de palabras utilizadas, la forma de suministrar las instrucciones o un inadecuado manejo del idioma por parte del examinador o del examinado. Tam-

bién comprende los cambios introducidos en el modo de administración de la prueba puesto que, frecuentemente, los manuales de tests incluyen instrucciones que no son adecuadas para la población evaluada. Los administradores del test, entonces, optan por "adaptar" esas instrucciones según su criterio personal, lo cual puede conducir a severas distorsiones en la interpretación de los resultados obtenidos. En la investigación mencionada de Fernández y Marcopulos (2004) se observó que algunas modificaciones en la metodología para cronometrar un test de velocidad producían puntuaciones medias muy diferentes en una muestra de individuos daneses y otra de individuos de la Argentina.

Sesgo de ítem

El sesgo del ítem se genera cuando éste último posee diferentes significados en las culturas consideradas. Ciertos grupos culturales pueden obtener puntajes significativamente distintos en un ítem determinado a pesar de obtener un puntaje total similar en el test. La deseabilidad social o la relevancia cultural, entre otros factores, pueden producir el sesgo de ítem. Tanzer (1995), por ejemplo, demostró que, aunque la estructura factorial de un test de autoconcepto académico era semejante en muestras de estudiantes australianos y singapurenses, existían diferencias sustanciales entre ambos colectivos cuando se comparaban los puntajes obtenidos en algunos ítems específicos.

7.3. La influencia del lenguaje

El lenguaje es otro factor que afecta el desempeño en la evaluación transcultural que utiliza tests psicológicos. Lau y Hoosain (1999) demostraron que los individuos que hablan chino rinden más que los sujetos que hablan japonés en una prueba de cálculo mental. Estos últimos, a su vez, superaron en su desempeño a las personas angloparlantes. Los autores pudieron demostrar que estas diferencias estaban relacionadas con la duración de los dígitos cuando son pronunciados, lo que a su vez está vinculado con la memoria de trabajo. Cuando les administraron una prueba de supresión articulatoria,[1] las diferencias entre estos tres grupos desaparecieron, lo que apoya la hipótesis de que la menor duración de los dígitos en el japonés que en el inglés les otorgaba ventaja a los japoneses y, a su vez, la menor duración de la pronunciación de los dígitos en chino comparado con el japonés e inglés, les daba ventaja a los chinos sobre los otros dos grupos.

Esta investigación demostró claramente cómo el idioma puede producir diferencias en el rendimiento en un test determinado y, por consiguiente, advierte sobre la inconveniencia de utilizar baremos extranjeros. En este caso específico, si los investigadores japoneses o chinos hubieran utilizado baremos elaborados en Inglaterra para evaluar el rendimiento de un individuo en un test de cálculo mental (algo que fácilmente podría ocurrir) es probable que los examinados hubiesen calificado dentro de rangos normales cuando en realidad algunos podrían tener un déficit.

7.4. Métodos de adaptación

Actualmente se reconoce que la adaptación de un test es un proceso mucho más complejo que la mera traducción a un idioma diferente. La traducción del inglés al español del siguiente ítem de un inventario de personalidad: "I wouldn't enjoy vacationing in Las Vegas" por "No disfrutaría tomando mis vacaciones en Las Vegas" es correcta. Sin embargo este ítem probablemente tendrá un significado distinto para personas que no viven en los Estados Unidos. Es evidente que una traducción correcta no asegura un significado unívoco de un ítem en diferentes culturas.

1. La capacidad de la memoria de trabajo puede estimarse a través de la amplitud de dígitos, es decir, la cantidad máxima de dígitos que puede recordar una persona en un ensayo. La supresión articulatoria se ejecuta para suprimir el componente fonológico de esta memoria a través de la repetición constante de una misma sílaba, por ejemplo "bah".

Van de Vijver y Leung (1997) establecieron tres niveles de adaptación de las pruebas psicológicas. El primero corresponde a la aplicación, esto es, la simple traducción de un test asumiendo a priori la equivalencia de constructo. Desafortunadamente, éste es el método más utilizado y, como se afirmó anteriormente, el problema es que la mera traducción de un test no garantiza la equivalencia entre ambas versiones (la original y la traducida).

La segunda opción es la adaptación. En este caso, a la traducción se agrega la transformación, adición o sustracción de algunos ítems de la escala original. En efecto, algunos ítems pueden cambiar su significado a través de las culturas y por lo tanto necesitan ser modificados o eliminados. Asimismo, los ítems nuevos que no existían en la versión original del test pueden representar mejor el constructo en la población en la cual se administrará la versión adaptada. En una investigación se descubrió que el nivel de dificultad original de algunos ítems pertenecientes a los subtests de Comprensión, Vocabulario e Información del WISC-III (Wechsler, 1994) no era apropiado para una muestra de estudiantes argentinos, por lo que se propuso un nuevo ordenamiento de los ítems (Baldo, 2001).

Finalmente, en la opción de ensamble (*assembly*) el instrumento original se modifica tan profundamente que prácticamente se transforma en un nuevo test. De allí el término ensamble (unión, integración), puesto que este proceso implica una integración del instrumento original con los nuevos elementos incorporados. Esto acontece cuando muchos de los ítems del test original son claramente inadecuados para representar el constructo a medir. Esta situación es común en los tests de denominación confrontacional, donde se presenta una lámina con dibujos de objetos que el examinado debe nombrar. Las láminas de estas pruebas son ordenadas desde objetos fáciles a difíciles de denominar. Como puede advertirse, este ordenamiento puede variar considerablemente de una cultura a otra. Este fue el caso de la adaptación argentina del Test de Denominación de Boston (Allegri y colaboradores, 1997). En la versión original, por ejemplo, la figura de una bellota está ubicada en el lugar número 32, mientras que en la versión argentina esa lámina se encuentra en el número 59 (el test consta de 60 láminas) y esta situación se presentaba en muchos ítems.

El ensamble también es recomendable cuando un constructo no está representado de forma adecuada por la versión original de un test que se desea usar en otra cultura. Los enfoques indigenistas[2] de la medición de la personalidad, por ejemplo, han promovido el desarrollo de tests distintos para medir rasgos de la personalidad no contemplados en las teorías desarrolladas en occidente, tales como la de los cinco grandes factores. Tal es el caso del Inventario Chino de Evaluación de la Personalidad (Ho, 1998), que mide dimensiones indigenistas de la personalidad tales como "armonía", junto a los otros factores reconocidos en occidente. En algunos casos, los límites entre "ensamble" y construcción de un nuevo test son poco nítidos, tal como puede apreciarse en este último ejemplo.

En el proceso de adaptación de un test hay una secuencia de actividades: en primer lugar es necesario traducir la versión original al idioma de la cultura meta del test; en segundo término se requiere introducir las modificaciones que sean necesarias a la nueva versión; y, finalmente, debemos verificar mediante un diseño de investigación apropiado que las dos versiones (original y adaptada) son equivalentes. A continuación revisaremos brevemente cada uno de estos pasos y concluiremos con un ejemplo concreto de adaptación de una prueba.

Técnicas de traducción de un test

El proceso de traducción es complejo e implica más que la traducción literal de las palabras escritas a un nuevo lenguaje. Existen dos métodos fundamentales: la traducción directa (*forward translation*) e inversa (*backward translation*).

En el método de traducción directa un traductor o, preferentemente, un grupo de traductores, traducen el test desde el idioma original al nuevo. Luego otro grupo de traductores juzga la

2. "Psicología indigenista es el estudio de la conducta humana y los procesos mentales dentro de un contexto cultural que descansa sobre los valores, conceptos, sistemas de creencia, y otros recursos del grupo étnico o cultural que se investiga" (Ho, 1998: 94).

equivalencia entre las dos versiones. De este modo pueden realizarse las correcciones pertinentes a las dificultades o errores identificados.

En el caso de la traducción inversa, el método más utilizado, un grupo de traductores realiza una traducción desde el idioma original al nuevo idioma; luego un segundo grupo de traductores toma el test traducido (en el nuevo idioma) y vuelve a traducirlo al idioma original. Seguidamente, se realizan las comparaciones entre la versión original y la versión retraducida al idioma original para determinar su equivalencia.

Ambos métodos poseen ventajas y desventajas que no serán analizadas en este texto introductorio. El lector interesado en profundizar esta problemática específica puede consultar el texto clásico de Hambleton (1994).

Diseños experimentales para verificar la equivalencia de tests e ítems

Una vez que se ha traducido adecuadamente el test, es necesario establecer si esta versión traducida es equivalente a la original. Para ello se requiere un diseño experimental y el análisis de los datos obtenidos a través de ese diseño. Hambleton (1994) indicó que existen básicamente tres métodos:

a) *Administración del test original y traducido a individuos bilingües.* En este caso se les administra ambas versiones del test (la original y la traducida al nuevo idioma) a personas que hablen ambos idiomas. Si, por ejemplo, quisiéramos adaptar un test de inteligencia desde el inglés al español, administraríamos la versión en inglés y la versión en español a individuos que hablen ambos idiomas, y luego verificaríamos la equivalencia de las puntuaciones de ambas formas. Este método posee ventajas y limitaciones. Entre las primeras se puede mencionar que permite controlar las diferencias de los participantes en el test (en su inteligencia, por ejemplo), puesto que ambas versiones del test son administradas a las mismas personas. Entre las desventajas, Hambleton señaló que este diseño está basado en la premisa de que los individuos son igualmente competentes en ambos idiomas, lo cual es difícil de sostener. Es probable, entonces, que puedan observarse diferencias entre los resultados de ambas versiones debido a una menor capacidad de algunas personas para entender los ítems en alguna de las dos lenguas. La segunda desventaja de este diseño es que no puede asegurarse que los bilingües posean el mismo nivel de competencia que la población general. Por el hecho de conocer otro idioma, es probable que se trate de personas con una mayor capacidad intelectual o mejor educación. Hambleton (1994.) también propuso una variación de este método, que conserva las mismas ventajas y desventajas pero que es más fácil de implementar, y consiste en administrar al azar una de las versiones del test (en inglés o en español, siguiendo nuestro ejemplo anterior) a cada participante bilingüe.

b) *Administración de la versión original del test y su traducción inversa a monolingües en el idioma original.* Siguiendo el ejemplo anterior, se les administraría la versión original del test de inteligencia en inglés y la versión obtenida por traducción inversa a personas cuyo idioma natal es el inglés. La equivalencia entre los ítems se determinaría comparando el desempeño de cada individuo en ambas versiones. Nuevamente, la ventaja de este diseño está en el control de las diferencias en las características de los participantes. Una limitación es que no permite obtener datos de la versión en el idioma meta del test (español, en este ejemplo). Otra desventaja reside en la posible falta de independencia entre los puntajes obtenidos, puesto que es probable que exista un efecto de aprendizaje luego de la administración de la primera versión de la prueba, especialmente si es la original. La administración al azar de una de las versiones en primer lugar puede atenuar la influencia de la variable "aprendizaje".

c) *Administración de la versión original a monolingües que hablan el idioma original y de la versión traducida a monolingües que hablan el idioma al que ha sido traducido el test.* Continuando con el ejemplo anterior, en este caso se

administraría la versión en inglés del test a sujetos cuya lengua materna es el inglés y se administraría la versión en español a personas cuyo idioma materno es el español. Una posible dificultad de este diseño reside en asumir que los sujetos de ambas muestras poseen una inteligencia (u otro rasgo medido) comparable. Sin embargo, Hambleton (1994) sugirió que este obstáculo puede subsanarse si los análisis son desarrollados en el contexto de la teoría de respuesta al ítem, que permite utilizar distintos conjuntos de ítems para obtener las mismas estimaciones del rasgo o aptitud.

Una vez obtenidos los datos por medio de estos diseños de investigación existen varias posibilidades de análisis estadístico. Básicamente, el análisis estará destinado a identificar que las propiedades psicométricas esenciales (confiabilidad, validez) del test original se mantengan en la versión adaptada. Para ello es necesario replicar esos estudios psicométricos del test original en la población meta de adaptación (consistencia interna, estructura interna mediante análisis factorial, estabilidad, evidencias de validez convergente-discriminante y relacionada con criterio, por ejemplo).

También es importante verificar la posible existencia de funcionamiento diferencial del ítem (FDI), es decir, corroborar que las propiedades psicométricas de los ítems (y no solamente de las puntuaciones totales del test) no varíen en las diferentes muestras. El FDI es una aplicación importante de la teoría de respuesta al ítem que revisaremos en el último capítulo.

En una investigación donde se compararon estudiantes australianos con estudiantes singapurenses a propósito del autoconcepto académico en lectura y matemática (Tanzer, 1995), pudo observarse que, a pesar de que la prueba mostraba la misma estructura factorial en ambos grupos, algunos ítems de la escala competencia/facilidad evidenciaban FDI. El autor conjeturó que esas diferencias entre ambas muestras podrían atribuirse a un factor de "modestia", una virtud deseable dentro de la cultura oriental que provocaba en los estudiantes singapurenses mayor resistencia a expresar una actitud de auto-confianza. En otro estudio, Galibert, Aguerri, Lozzia y Abal (2006) analizaron

el funcionamiento diferencial de los ítems de una escala construida para medir un constructo semejante a Responsabilidad correspondiente al modelo de los cinco factores de personalidad. En esa investigación, en una muestra de estudiantes universitarios de Psicología, encontraron que el ítem "Aunque me sienta cansado finalizo la tarea que me impuse" evidencia posible interacción entre el constructo medido y el sexo.

El FDI se detecta con varios estadísticos que, esencialmente, consisten en refutar o confirmar la hipótesis nula que indica su ausencia. El método más utilizado en la actualidad es el de Mantel y Haenszel (1959) que fuera adaptado para el FDI por Holland y Thayer (1988).

Ejemplo de adaptación de un test

La Comisión Internacional de Tests (International Test Commission) ha elaborado 22 pautas para adaptar tests de una cultura a otra (Hambleton, 1994). Esta serie de recomendaciones comprende cuatro secciones: contexto, desarrollo y adaptación del instrumento, administración e interpretación.

a) Contexto: Por ejemplo, "los efectos de las diferencias culturales que no sean relevantes para los objetivos centrales del estudio deberían minimizarse en la medida de lo posible".
b) Desarrollo y adaptación del instrumento: Por ejemplo, "los constructores/editores de tests deberían asegurar que el proceso de adaptación tiene en cuenta las diferencias lingüísticas y culturales entre las poblaciones a las que se dirigen las versiones adaptadas del test".
c) Administración: Por ejemplo, "los constructores y administradores de tests deberían tratar de prever los problemas de comprensión que pueden presentarse en la administración del test y tomar las medidas oportunas para evitarlos mediante la preparación de materiales e instrucciones adecuados".
d) Interpretación: Por ejemplo, "cuando se adapta un test para utilizarlo en otra población, debe facilitarse la docu-

mentación sobre los cambios introducidos, así como los datos que permitan verificar la equivalencia entre las versiones".

El establecimiento de estos estándares constituye un importante marco conceptual de referencia para el arduo trabajo de adaptar tests. Aunque estas pautas no se abordarán en detalle aquí, se analiza a continuación un ejemplo de aplicación de las mismas al proceso de adaptación de un test.

En un estudio reciente, El-Hassan y Jammal (2005) adaptaron a la cultura del Líbano el Test de Comprensión Auditiva del Lenguaje - Revisado (Carrow-Woolfolk, 1985). Este test, construido en los Estados Unidos, consta de tres partes que miden la comprensión del significado común y literal de las palabras, así como la comprensión de morfemas gramaticales y el significado de oraciones complejas. Es apropiado para niños de 3 a 10 años y sus ítems consisten en seleccionar entre varias figuras aquella que mejor representa el estímulo verbal provisto por el examinador. Estos investigadores realizaron una traducción directa desde el inglés al árabe y propusieron las siguientes modificaciones en los ítems para adaptar los materiales:

a) tradujeron al árabe las palabras y frases que aparecían en algunas láminas;
b) cambiaron la figura del billete de un dólar por un billete de mil libras libanesas en el ítem 34;
c) la tercera figura del ítem 3 en la sección II se modificó para mostrar la imagen de un campesino libanés en lugar de un campesino occidental;
d) se modificó la figura que representaba una pasta dental occidental, introduciendo una figura similar a la pasta dental que utilizan los niños en el Líbano;
e) en un ítem que incluía figuras de guantes de béisbol, las mismas se reemplazaron por imágenes de guantes comunes.

Se administró entonces la versión adaptada a una muestra piloto de 50 niños. De acuerdo a los resultados obtenidos, se realizó un reordenamiento de los ítems según la frecuencia de dificultad en la versión árabe. De esta manera, algunos ítems cambiaron radicalmente su ubicación respecto de la versión original en inglés. Por ejemplo, el ítem 22 ("cuatro") pasó a ser el número 15 en la versión en árabe y el ítem 8 ("el gato está entre las sillas") pasó a la posición 14. Los investigadores observaron que los ítems que incluían preposiciones y verbos parecían ser más difíciles que aquellos que incluían sustantivos y descripciones de cantidad.

Finalmente, se realizaron estudios psicométricos esenciales (consistencia interna, validez) y los investigadores estandarizaron el test para utilizarlo con niños libaneses. Éste es un claro ejemplo de un test desarrollado en inglés y adaptado rigurosamente a una lengua (y una cultura) tan disímil como la árabe.

Tercera Parte

TEORÍA DE LOS TESTS

8
TEORÍA CLÁSICA DE LOS TESTS

Silvia Tornimbeni - Edgardo Pérez

Las mediciones a través de tests psicológicos se realizan sobre individuos en un momento determinado de sus vidas, en circunstancias particulares (administradores, ambientes), y utilizando una muestra específica de comportamiento que se supone representativa de un dominio o constructo generalmente inobservable. Estas limitaciones implican que el resultado de la medición (la puntuación observada en el test) es sólo una estimación aproximada de la puntuación verdadera del individuo en el constructo de interés (inteligencia, por ejemplo).

Además de esas características particulares de los tests psicológicos debe considerarse que toda medición, cualquiera sea el campo científico en que se implemente, conlleva errores. Aun en las ciencias naturales, donde existen instrumentos más precisos de medición que en las ciencias sociales, existe esta posibilidad de error. La lectura o registro del dato puede variar de un evaluador a otro, o en distintas mediciones realizadas a un mismo individuo. La Psicometría hace uso de modelos formales que le permiten lograr esa estimación de la puntuación verdadera de un individuo en un rasgo con el menor error posible.

Como ya expresamos en el apartado histórico precedente, la teoría clásica de los tests (TCT) surge del modelo lineal de medición formulado por Spearman (1927) y fue consolidada mediante los desarrollos teóricos y metodológicos de Thurstone (1935) y Gulliksen (1950), entre otros investigadores. Se considera que los errores de medición varían en torno a algún valor verdadero (Martínez Arias, 1995). Gauss y otros matemáticos del siglo XVIII derivaron la función de distribución normal con

media cero, que permitió describir la distribución de los errores de medida. Éste es uno de los antecedentes históricos más importantes de la TCT.

Como afirma Muñiz (2001), el modelo lineal de medición sobre el que se asienta la TCT es sencillo, robusto y parsimonioso, y satisface la mayor parte de las necesidades de los profesionales de la medición psicológica, tanto en lo relativo a la confiabilidad de las mediciones (estimación del error) como a la validez (inferencias hechas a partir de los tests).

Esta teoría se basa en supuestos débiles pero generales, vale decir, adaptables a distintas situaciones. La hipótesis fundamental de la teoría clásica es que la puntuación observada de una persona en un test es una función lineal de dos componentes: su puntaje verdadero (inobservable) y el error de medición implícito en la prueba. La TCT es un modelo de puntuación verdadera como valor esperado. El término esperado alude a que la puntuación verdadera es un concepto matemático, probabilístico. La puntuación verdadera de un sujeto en un test estaría dada por el promedio aritmético de las puntuaciones empíricas obtenidas en infinitas aplicaciones (Muñiz, 2001). La puntuación verdadera es constante, pero las puntuaciones observadas y los errores son aleatorios. Esta última afirmación equivale a decir que no podemos predecir con certeza cuáles serán las puntuaciones observadas y los errores en las diferentes aplicaciones de un mismo test a una persona o grupo de personas.

Formalmente, estas últimas consideraciones pueden expresarse como:

a) para cada individuo*, su puntuación observada es igual a la puntuación verdadera más el error*: $O_i = V_i + E_i$
b) el error es igual a la puntuación observada menos la verdadera: Ei = Oi -Vi

Los corolarios que se desprenden de los supuestos anteriores son:

1. La media de los errores (μe) tiende a cero cuando el número de mediciones tiende al infinito:

$$\mu e = 0$$

2. La media de las puntuaciones verdaderas (Jv) es igual a la media de las puntuaciones observadas (Jo) cuando N tiende al infinito:

$$\mu v = \mu o$$

Debe aclararse que los dos corolarios anteriores parten del supuesto de que los errores aleatorios y las puntuaciones observadas se distribuyen de manera normal. Para un repaso del concepto de forma de distribución y distribución normal, véase el capítulo 5 sobre interpretación de las puntuaciones.

3. Si extraemos una muestra aleatoria de una población y le administramos un test: a) las puntuaciones observadas tendrán una distribución normal y b) los límites que encierran a la media poblacional (μ), con una probabilidad del 95%, estarán dados por:

$$\mu_o \pm 1{,}96\ \sigma$$

La media de puntuaciones en el test de la población se ubicará con una probabilidad del 95% dentro del intervalo comprendido entre 1,96 desviación estándar a la izquierda y a la derecha de la media de las puntuaciones de la muestra.

4. La varianza de las puntuaciones observadas (σ2o) es igual a la varianza de las puntuaciones verdaderas (σ2v) (la varianza del atributo que se está midiendo) más la varianza del error de medición (σ2e):

$$\sigma^2_o = \sigma^2_v + \sigma^2_e$$

5. La correlación entre las puntuaciones observadas de dos formas paralelas de un mismo test (rxx') es igual a la razón entre la varianza de las puntuaciones verdaderas (σ2v) sobre la varianza de las puntuaciones observadas (σ2o):

$$r_{xx'} = \frac{\sigma_v^2}{\sigma_o^2}$$

Deduciendo de lo anterior: la varianza de las puntuaciones verdaderas es igual al producto de la varianza de las puntuaciones observadas por la correlación de las medidas paralelas. Esta deducción es importante, puesto que la correlación entre dos formas paralelas de un test (un dato observable) nos permite estimar la varianza de las puntuaciones verdaderas, que es inobservable. Derivando la fórmula anterior obtenemos el corolario que sigue.

6. El error estándar de medición (desviación estándar del error) es igual a:

$$EEM = s_t . \sqrt{1 - r_{xx}}$$

Donde:

s_t = desviación estándar de las puntuaciones del test en la muestra de estandarización
r_{xx} = coeficiente de correlación test-retest

Para comprender este corolario debe considerarse que la desviación estándar (raíz cuadrada de la varianza) de las puntuaciones observadas en repetidas mediciones a un individuo es en sí misma un índice de error. En efecto, si la desviación estándar de esas repetidas mediciones fuera igual a cero no habría error de medición. Nunca tenemos un número suficiente de observaciones para calcular el error estándar de medición de la distribución de los puntajes observados. Sin embargo, como la correlación test-retest (véase capítulo 3 de confiabilidad) se obtiene de dos administraciones del test a una misma muestra, cuanto más alta es la correlación menor debería ser el error de medición (Kline, 2000). Este supuesto de la igualdad de las varianzas de error a lo largo de todo el continuo de puntuaciones se denomina homoscedasticidad y es uno de los supuestos más cuestionables de la teoría clásica de los tests, como veremos más adelante.

De estos supuestos, corolarios y deducciones puede concluirse que (Murat, 1985):

1. La puntuación verdadera de un individuo es una puntuación "límite", un punto en un intervalo de la distribución de puntuaciones observadas.
2. Cuanto más alta sea la confiabilidad de un test, menor será ese intervalo, y la puntuación observada se ubicará más próxima a la puntuación verdadera.
3. Como la desviación estándar no puede modificarse, deben disminuirse los errores de medida.

Como argumenta Muñiz (2001), la aplicación de este modelo presenta algunos problemas de medición que pueden enunciarse del siguiente modo:

a. La TCT suministra un coeficiente de confiabilidad integral (para todo el test), pero los datos de investigaciones empíricas demuestran que la precisión de los tests también depende del nivel de desempeño del individuo en el constructo medido. En síntesis, los tests no miden con la misma precisión a todos los individuos.
b. Las puntuaciones no son invariantes respecto del instrumento utilizado. Así, por ejemplo, si se utilizan tres tests distintos para medir la inteligencia de tres personas, estrictamente no se puede comparar el nivel de inteligencia de las mismas.
c. En el marco de la TCT, las propiedades métricas (confiabilidad, validez) de los instrumentos de medida no son invariantes respecto de los individuos evaluados para estimarlas. Por consiguiente, las propiedades de los ítems de un test (dificultad, por ejemplo) también dependen de las muestras de investigación.

Cronbach y Gleser (1972), paralelamente a los mejores desarrollos de la TCT (como los coeficientes de consistencia interna alfa y Kuder-Richardson), expusieron la teoría de la generalizabilidad (TG), que puede considerarse una extensión de la teoría clásica de los tests para superar algunas de sus limitaciones (Nunnally y Bernstein, 1995).

La teoría de la generalizabilidad postula que el concepto de confiabilidad en el contexto de la TCT es muy limitado, pues no tiene en cuenta todas las posibles fuentes de error. Como vimos antes, el error estándar de medición se estima a partir del coeficiente de correlación test-retest y, por consiguiente, sólo controla una de las fuentes posibles de error aleatorio (los cambios aleatorios de los individuos examinados). Los diseños de la TCT son limitados para alcanzar un control simultáneo de varias fuentes de error. Debido a que las mediciones en ciencias del comportamiento pueden incluir múltiples fuentes de error, la TG suministra herramientas más sólidas de control. En efecto, los diseños de análisis de varianza utilizados por la TG permiten discriminar varios componentes de varianza de error en un único análisis.

Cuando se administra un test, en el puntaje observado de un individuo operan diversas influencias además de las variaciones de su aptitud o rasgo verdadero. En este sentido, pueden mencionarse el ambiente de administración del test, la particular muestra de ítems del mismo, las instrucciones de respuesta dadas por el psicólogo, entre otras. Cada una de estas variaciones es una fuente de varianza de error y el puntaje observado de las personas puede cambiar en función de las mismas.

En la TG el coeficiente de confiabilidad se denomina coeficiente de generalizabilidad. Al igual que en la TCT, el coeficiente de generalizabilidad refleja la proporción de variabilidad de las puntuaciones observadas que puede atribuirse a la variación sistemática de las puntuaciones verdaderas (universo, en la terminología de la TG). El concepto de generalizabilidad se refiere a cuán precisamente podemos generalizar el puntaje observado al puntaje promedio que la persona debería recibir considerando todas las posibles condiciones de administración de un test.

Las diferentes fuentes de variación aleatoria en las puntuaciones del test se denominan facetas en la terminología de la TG. Este término fue introducido para designar cada una de las características de la situación de medición que pueden cambiar de un momento a otro y, por consiguiente, hacer variar los resultados obtenidos. Así, la faceta "ítems" sería la variedad particular de ítems que los usuarios de un test pueden encontrar en el mismo.

El análisis de varianza (ANOVA) proporciona las herramientas estadísticas para realizar un estudio de generalizabilidad. Debe tenerse en cuenta que el ANOVA permite estudiar de manera simultánea el efecto de varias variables independientes sobre una variable dependiente, así como las interacciones, o sea el efecto que genera la combinación de dos o más variables además de sus efectos independientes (Hogan, 2004).

En el lenguaje del ANOVA, las facetas serían los factores, y sus efectos combinados con otras facetas constituirían las interacciones. A los factores del ANOVA se les ha dado el nombre de facetas (en el contexto de la TG) para evitar la confusión con los factores del análisis factorial (véase cap. 6 sobre construcción de tests).

En los diseños de una faceta (dificultad de los ítems, por ejemplo) habría cuatro fuentes de variación, que en un test de rendimiento en matemática podrían ser:

- diferencias individuales de los estudiantes en ese dominio;
- diferencias de dificultad de los ítems;
- interacciones de las personas con los ítems;
- errores aleatorios y fuentes de error no identificadas.

9
TEORÍA DE RESPUESTA AL ÍTEM

Nuria Cortada de Kohan - Marcos Cupani

9.1. Introducción

Un test no es un instrumento de medición como un metro o un velocímetro, los cuales posibilitan mediciones directas de una propiedad (longitud o velocidad, por ejemplo) de un objeto en una escala numérica. En realidad, un test debe considerarse como una serie de pequeños experimentos, donde el profesional registra y califica un conjunto de respuestas de un individuo. La calificación de estas respuestas a los ítems de un test no constituyen medidas directas sino que proporcionan los datos a partir de los cuales inferimos el nivel de un individuo en un dominio (área específica de conocimiento) o en un constructo o rasgo inobservable (latente).

Como en cualquier experimento, en los tests se presenta el problema de controlar en las respuestas el error de medida. Este error se ocasiona porque en el test no operan solamente las variables independientes (constructos medidos), sino también otras variables (revisadas en el capítulo 3 sobre confiabilidad) que pueden influir en las respuestas. El control de estas variables de error se efectúa mediante tres procedimientos fundamentales: 1) la estandarización, 2) la aleatorización, y 3) el ajuste estadístico. Esta última es la forma específica de control del error en el contexto de la teoría de respuesta al ítem.

Una de las principales limitaciones que enfrenta la teoría clásica de los tests (TCT) es la relativa al control de los errores implícitos en cualquier medición psicológica. Desde los primeros trabajos de Spearman (1927) hasta la actualidad han sido

muchos los esfuerzos para desarrollar nuevas teorías y técnicas que permitieran su identificación y estimación. Algunos de estas teorías pueden considerarse extensiones y refinamientos de la TCT (por ejemplo, la teoría de la generalizabilidad); sin embargo, entre los años sesenta y la década de los ochenta, surge y se consolida un nuevo modelo dentro del campo de la teoría de los tests, la teoría de respuesta al ítem (TRI) que permite superar algunos de los inconvenientes de la teoría clásica de los tests (Van der Linden y Hambleton, 1997).

Los diferentes desarrollos de la TRI se deben fundamentalmente a varias líneas de investigación independientes. Una de ellas se desprende del texto clásico de Lord y Novick, *Statistical theories of mental test scores* [Teorías Estadísticas de los Puntajes de los Tests Mentales], de 1968, comparable en su impacto al texto de Gulliksen *Theories of mental tests* de los años cincuenta. Por otra parte, el matemático danés Georg Rasch desarrolló una serie de modelos teóricos TRI para tests de lectura y para uso militar en la década del sesenta. Finalmente, los avances en la informática posibilitaron la extensión y difusión de los complejos cálculos requeridos en los diferentes procedimientos TRI.

En la actualidad, varios tests han sido construidos con métodos de la TRI, tales como la batería psicoeducacional Woodcock-Johnson, el test de aptitudes académicas (SAT), la batería de aptitudes vocacionales de las fuerzas armadas (ASVAB) y la actual versión del test de inteligencia Stanford-Binet, todos mencionados en las páginas precedentes de este texto (Embretson y Reise, 2000).

Como mencionamos repetidamente en este libro, la TCT se apoya en un supuesto fundamental (O=V+E), según el cual la puntuación empírica (observada) de una persona en un test está integrada por su puntuación verdadera en el rasgo medido (inobservable) y el error de medición. A partir de este supuesto, y otros adicionales, se deducen postulados que fundamentan las estimaciones empíricas de la confiabilidad y validez de los tests, así como los indicadores de las propiedades psicométricas de sus ítems (índices de discriminación o dificultad, por ejemplo).

La TCT asume que la aptitud de un examinado se define en términos de un test particular y en el número de respuestas correctas (en el caso de tests de ejecución máxima) o clave (en los tests de respuesta típica) emitidas ante el mismo. Este supuesto es muy controversial, puesto que si un test es "difícil" (sus ítems son dificultosos) el examinado obtendrá un puntaje que expresará "poca aptitud", pero si un test que mide el mismo rasgo es "fácil" (sus ítems son más fáciles) el mismo examinado parecerá tener "mucha aptitud", aunque posea el mismo nivel de aptitud o rasgo. Consideremos dos individuos que responden correctamente el 50% de los ítems de dos tests de inteligencia que difieren en dificultad, ¿podrían ser calificados como igualmente inteligentes? Claramente, la respuesta es no. Por otra parte, dos personas pueden responder correctamente al mismo número de ítems en un test (y por consiguiente obtener el mismo puntaje total) pero diferir en su aptitud. La persona que hubiese acertado los ítems más difíciles tendría un puntaje más elevado en el rasgo latente que el individuo que hubiese acertado el mismo número de ítems pero más fáciles.

Todas las fuentes potenciales de variabilidad (diferentes del nivel real en el rasgo) de los puntajes se asumen constantes en la TCT o bien tienen un efecto no sistemático (aleatorio). Sin embargo, como afirmamos antes, los puntajes de dos tests construidos para medir el mismo rasgo suelen ser distintos aunque se hayan estandarizado cuidadosamente. Esto se debe a que cada test tiene su propio conjunto de ítems y cada ítem tiene diferentes propiedades. En efecto, las propiedades de los ítems son variables de error que evaden la estandarización del test en la TCT.

La teoría clásica ha sido muy útil y aún continúa vigente, coexistiendo con la TRI, pero se ha señalado que posee algunas limitaciones importantes (Baker, 2001; Hambleton y Rogers, 1991; Muñiz, 2000; Embretson y Reise, 2000):

a) Las propiedades psicométricas de los ítems (por ejemplo, los índices de dificultad y de discriminación) y del test en su conjunto (confiabilidad y validez) dependen de las características de la muestra de estandarización del test.
b) El rendimiento de las personas en dos tests diferentes que miden un mismo rasgo no es estrictamente comparable (por ejemplo, si poseen ítems de diferente dificultad).

c) La TCT asume que la precisión o confiabilidad (y el error estándar de medición) con la que se mide el nivel de rasgo en un test es la misma para todos los examinados, y éste es un supuesto que generalmente no se verifica empíricamente.

La TRI permite superar estas limitaciones de la TCT mediante unos supuestos más fuertes y restrictivos (las condiciones que deben cumplirse para ser aplicada) y con una metodología más sofisticada. Sin embargo, no debe olvidarse que el objetivo básico de ambas teorías es el mismo: obtener la puntuación que corresponde a una persona en un rasgo latente o dominio de conocimiento. En este sentido, ambas teorías consideran que el puntaje de cada individuo en un test se asocia a un parámetro individual inferible, que en la teoría de la respuesta al ítem se denomina theta, simbolizado por la letra griega θ, equivalente al puntaje verdadero (V) en la teoría clásica.

Por otro lado, ninguna de las dos teorías permite alcanzar un nivel proporcional de medición, una de las objeciones realizadas a la medición psicológica en su conjunto (y que revisamos en el capítulo inicial de este libro). Adicionalmente, algunos autores destacan que, pese a las ventajas teóricas de la TRI, en la práctica las correlaciones entre tests semejantes construidos según los dos modelos teóricos principales (TCT y TRI) son elevadas (Kline, 2000; Nunnally, 1991). No obstante, la TRI posibilita aplicaciones importantes no factibles con los métodos de la TCT, tales como la generación de bancos de ítems para construir tests de rendimiento, tests adaptativos computarizados o el análisis del funcionamiento diferencial del ítem, que revisaremos más adelante.

La TRI debe su nombre a que se interesa más en las propiedades de los ítems individuales que en las propiedades globales del test, como hace la TCT (Abad, Garrido, Olea y Ponsoda, 2006), y utiliza el ajuste estadístico (Van der Linden y Hambleton, 1997) como método de control del error de medición. Este método requiere explicitar los parámetros de la aptitud que nos interesa medir así como aquellos que corresponden a las propiedades de los ítems según un modelo determinado. Si este modelo se sostiene (se ajusta a los datos reales) y los parámetros de los ítems se conocen, puede ser usado para obtener mediciones de la aptitud que estén libres de las propiedades de error de los ítems del test. Recordemos que "parámetro" es el índice de una variable (media, desviación estándar, varianza) en una población. Por el contrario, hablamos de estadísticos para referirnos a esos índices cuando son estimados en una muestra. Generalmente, los parámetros son desconocidos y se estiman a partir de los estadísticos.

Otra diferencia esencial entre la teoría clásica de los tests y los diversos modelos de la teoría de la respuesta al ítem es que la calificación de un test, en el contexto de la TCT, estima el nivel de un atributo (aptitud, por ejemplo) como la sumatoria de las repuestas a los ítems de un test (es decir, como una combinación lineal), mientras que la TRI utiliza el patrón probabilístico de respuesta, mediante funciones matemáticas del tipo de los modelos logísticos de 1, 2 ó 3 parámetros (Santisteban, 1990).

La función logística define una familia de curvas teóricas (entre ellas, la curva característica del ítem); fue derivada en 1841 en las ciencias biológicas y se utiliza en modelos explicativos del crecimiento de plantas y animales desde su nacimiento hasta la madurez. Esta función es muy semejante a la función normal acumulada y forma una curva en forma de S con valores de 0 a 1 en la ordenada (valores de probabilidad) y valores correspondientes al rasgo medido en la abscisa, como veremos más adelante.

La escala de medición más popular utilizada en los modelos TRI (para los parámetros de los ítems y del rasgo) es asimilable, mediante una transformación, a las puntuaciones estándar que vimos en el capítulo 5 sobre interpretación de puntuaciones, con 0 como punto medio de habilidad (rasgo) y las unidades de medida expresadas en desviaciones estándar, en un rango de - 3 a + 3. Por consiguiente, estas escalas TRI poseen las propiedades métricas de una escala de intervalo (véase capítulo de Fundamentos de la Medición Psicológica), nivel de medición que frecuentemente no alcanzan las escalas de medición construidas con los métodos de la TCT.

Los modelos TRI son funciones matemáticas que relacionan las probabilidades de responder acertadamente a un ítem con la aptitud general del individuo. Su origen no es reciente (los pri-

meros se formularon en la década del sesenta) pero, dada la complejidad de los cálculos requeridos, comenzaron a difundirse y a utilizarse con la creación de programas de computación específicos, tales como LOGIST, MULTILOG o BILOG. Puesto que la TRI es un modelo probabilístico, revisaremos algunos conceptos básicos del cálculo de probabilidades que nos facilitarán la comprensión de los conceptos desarrollados posteriormente.

Recordemos que la probabilidad es "la frecuencia relativa con que esperamos que suceda un determinado evento o resultado" (Aron y Aron, 2001: 157). La frecuencia indica cuántas veces ocurre un hecho determinado, y la frecuencia relativa es la cantidad de veces que ese hecho sucede en relación con la cantidad de veces que podría haber ocurrido. Por consiguiente, la probabilidad (frecuencia relativa) de un evento se obtiene mediante la razón entre la cantidad de veces que ocurre un evento y la cantidad de veces que podría haber sucedido. Si arrojamos una moneda una vez, con dos resultados posibles (cara o cruz) existe una probabilidad de 0,5 (1/2) de obtener cara. Si tiramos un dado (6 caras), la probabilidad de sacar un 3 (o cualquier otro resultado) es 0,17 (1/6), así como la probabilidad de obtener un 3, un 2 o un 1 es 0,5 (3/6). Las probabilidades no pueden ser menores a 0 ni mayores que 1 y, expresadas en porcentajes, van del 0% al 100%. La probabilidad de acierto se simboliza con la letra P y la probabilidad de error con la letra Q, que es igual a 1 – P.

En síntesis, la TRI suministra una fundamentación probabilística al problema de la medición de constructos inobservables (latentes), considerando al ítem como unidad básica de medida en lugar del puntaje total del test como en la TCT.

9.2. Los postulados y supuestos de la TRI

Cualquier modelo TRI establece una relación matemática entre la probabilidad de emitir una determinada respuesta a un ítem, las características del individuo (su nivel en uno o más rasgos) y las propiedades del ítem (su dificultad o discriminación, por ejemplo). Cuando se asume y se comprueba que el rendimiento en un ítem depende de un único rasgo latente (comprensión verbal, por ejemplo) se habla de modelos unidimensionales; cuando el desempeño en un ítem depende de dos o más rasgos se utilizan modelos multidimensionales. Por otra parte, si el modo de calificación de las respuestas es dicotómico, normalmente el que corresponde a ítems donde existen aciertos y errores, se usan modelos dicotómicos; si, en cambio, el test posee tres o más categorías de respuesta (en ítems de tests de personalidad, por ejemplo) se emplean modelos politómicos. Un modelo que se usa en las escalas *likert* es el de respuesta graduada (Samejima, 1973), una extensión del modelo de dos parámetros que comentamos más abajo. No obstante, los modelos dicotómicos (Rasch, por ejemplo) pueden ser adaptados para ser empleados en escalas *likert*. Ahora veamos cuáles son los postulados y supuestos de esta teoría.

La TRI se basa en tres postulados (Hambleton y Rogers, 1991):

a) El desempeño de un examinado en los ítems de un test puede explicarse a partir de una serie de factores inobservables, denominados rasgos, aptitudes o rasgos latentes (θ). Este postulado no se diferencia esencialmente de la TCT, salvo en la denominación del constructo latente en cada teoría (theta o puntuación verdadera).
b) La relación entre la respuesta al ítem y los rasgos subyacentes puede describirse con una función monótona denominada "curva característica del ítem" (CCI). Esta función especifica que en la medida en que aumenta el nivel del rasgo medido, la probabilidad de responder correctamente a un ítem aumenta también (excepto en ítems con discriminación negativa). En matemática, la expresión función o aplicación del conjunto A en B significa que cualquier elemento del conjunto A se asocia solamente con un elemento del conjunto B. La función monótona es aquella que conserva el orden entre conjuntos ordenados, pudiendo ser creciente o decreciente.
c) Tanto las estimaciones del rasgo latente (θ) obtenidas con ítems diferentes como las estimaciones de los parámetros

de los ítems obtenidos en distintas muestras serán iguales. Esta última propiedad de invarianza tiene, entonces, un doble sentido: invarianza de los ítems respecto a posibles diferentes distribuciones de la habilidad o del rasgo, e invarianza de la habilidad medida a partir de diferentes conjuntos de ítems. Los parámetros de los ítems deberán ser los mismos, tanto si éstos se han aplicado a un grupo de personas con elevados niveles de rasgo o a un grupo con niveles bajos. Esto implica que el nivel de habilidad de una persona puede ser obtenido a partir de conjuntos de ítems distintos, facilitando muchas aplicaciones importantes de la TRI, tales como los tests adaptativos computarizados y los bancos de ítems (Hambleton y Rogers, 1991; Muñiz, 2001).

Con respecto a los supuestos de la TRI, podemos destacar dos fundamentales. El primer supuesto es la *unidimensionalidad* del rasgo latente, por el cual debe demostrarse que los ítems de un test miden sólo una aptitud o rasgo (Cortada de Kohan, 1998). En otras palabras, el rendimiento en un ítem dependerá del nivel de la persona en un solo rasgo o dimensión. Por ejemplo, en un test que mida vocabulario en inglés, la probabilidad de emitir una respuesta correcta a sus ítems dependerá exclusivamente del nivel de esa habilidad en el examinado y de los parámetros de los ítems (dificultad, discriminación, probabilidad de adivinar la respuesta correcta). El supuesto excluye la probabilidad de que la respuesta correcta a los ítems de ese test varíe en función de otros rasgos relacionados, tales como el nivel de habilidad para el inglés hablado o el conocimiento de la gramática de esa lengua (Abad y colaboradores, 2006).

Sin embargo, este supuesto nunca se cumple totalmente porque el rendimiento en un test es también afectado por variables cognitivas y de personalidad, tales como motivación y ansiedad, entre otras. Por consiguiente, en la práctica, es una cuestión de grado y no puede afirmarse categóricamente que un conjunto de ítems es unidimensional. El grado de unidimensionalidad de un test puede evaluarse mediante técnicas de análisis factorial exploratorio (véase cap. 6, "Construcción de tests") aplicadas a la matriz de correlaciones entre los ítems (Lord y Novick, 1968),

mediante diferentes procedimientos, como una explicación aproximada del 30% de la varianza por el primer factor o el empleo del análisis paralelo (véase ese mismo capítulo).

Aparentemente, no habría diferencias entre la TCT y la TRI en relación con el supuesto de unidimesionalidad. Sin embargo, el análisis factorial dentro del modelo general lineal (utilizado en la TCT) asume variables continuas y distribución normal de las puntuaciones, supuestos que frecuentemente no se respetan y pueden conducir a una sobre o subestimación de los factores latentes. Además, en el caso de ítems dicotómicos, el empleo de la matriz de correlaciones tetracóricas no resuelve estas dificultades. Una solución ideal en la TRI es el empleo del análisis factorial no-lineal (Embretson y Reise, 2000), que no requiere supuestos tan exigentes como su contraparte lineal. Existen programas específicos para aplicar el análisis factorial a un test TRI, y que tienen en cuenta no solamente la correlación entre los ítems y los factores (pesos factoriales) sino también los parámetros de los ítems (dificultad, discriminación).

El segundo supuesto se refiere a la *independencia local*, que se verifica cuando la respuesta de un individuo a un ítem no depende de la respuesta que haya dado a los otros ítems del test. Este supuesto se deriva del anterior (unidimensionalidad) puesto que la respuesta a un ítem sólo es explicada por el nivel de theta y los parámetros de los ítems, y no está influida por el orden de presentación de los ítems u otras variables semejantes. Matemáticamente, este supuesto significa que la probabilidad (P) de responder correctamente a un conjunto de ítems es igual al producto de las probabilidades de responder correctamente a cada uno de esos ítems individualmente. En efecto, el principio de independencia local se basa en la regla de multiplicación de las probabilidades que permite calcular la probabilidad de obtener más de un resultado independiente. Obtener cara o cruz en un tiro de moneda es independiente de obtener cara o cruz en un segundo tiro. En este caso, si tiramos dos veces la moneda, la probabilidad de obtener cara en los dos tiros es 0,25 (0,5 por 0,5). Del mismo modo, si alguien tiene una probabilidad de 0,5 de responder correctamente a cada uno de dos ítems, la probabilidad de que el individuo responda correctamente a ambos ítems es igual a 0,25 (0,5 x 0,5).

Consideremos, por ejemplo, un test que comprenda dos ítems, con una probabilidad de que un individuo acierte el primero (P1) = 0,4 y de que acierte el segundo (P2) = 0,8. Esta probabilidad depende, naturalmente, del nivel que la persona tenga en el rasgo. Del mismo modo, tengamos en cuenta que la probabilidad de error (Q) es igual a 1 - P. El principio de independencia local establece que la probabilidad de que un individuo acierte los dos ítems del ejemplo viene dada por: (P1) (P2) = (0,4) (0,8) = 0,32. La probabilidad de acertar el primero y fallar el segundo sería (Q2 = 1 – P2) = 1 – 0,8 = 0,2, por lo tanto (P1). (Q2) = (0,4) (0,2) = 0,08. La probabilidad de que erre el primero y acierte el segundo será (Q1). (P2) = (0,6) (0,8) = 0,48, y la de que su respuesta sea incorrecta en ambos ítems será (Q1) (Q2) = (0,6) (0,2) = 0,12. Por consiguiente, si 100 personas con idéntico nivel de rasgo que ese individuo (habilidad, actitud, etc.) responden ese test, esperaríamos aproximadamente los resultados que aparecen en la tabla que sigue:

Tabla 9.1. Frecuencias del patrón de respuestas a dos ítems

Item 1	Ítem 2	Número de personas
1	1	32
1	0	8
0	1	48
0	0	12
		100

1 = acierto; 0 = error

Si correlacionamos las 100 respuestas al primer ítem con las 100 respuestas al segundo, la correlación de Pearson será 0. Este hecho sugiere un procedimiento para contrastar si el supuesto de independencia local se cumple. El mismo consiste en obtener la matriz de correlaciones entre los ítems, pero no en la muestra completa, sino en submuestras que sean homogéneas en el nivel de habilidad de sus miembros. En tales submuestras, si el supuesto se cumple, los ítems no deberían correlacionarse entre sí. Por consiguiente, en un grupo homogéneo de habilidad o aptitud las correlaciones entre los ítems deberían ser cercanas a 0 si este supuesto se cumple (Hambleton y Rogers, 1991).

De las consideraciones precedentes se desprende que la respuesta a un ítem no se correlaciona con la respuesta a otro cuando el nivel del rasgo es controlado. Es decir, los ítems de un test pueden estar altamente intercorrelacionados en la muestra total, pero si el nivel del rasgo es controlado (por ejemplo, si se examina a un subgrupo de habilidad homogénea, con niveles iguales de rasgo) el supuesto de independencia local se verifica si no existen correlaciones entre los ítems. En la práctica, este supuesto se viola cuando las respuestas a los diferentes ítems de un test están vinculadas, por ejemplo, cuando la respuesta a un ítem proporciona información relevante para responder a otro ítem del test (Embretson y Reise, 2000).

La independencia local es el supuesto fundamental y distintivo de la TRI y, recapitulando, significa que el nivel de theta (aptitud o rasgo latente del individuo) debería explicar totalmente la variabilidad de respuesta al ítem (Kline, 2000).

9.3. Curva característica del ítem y modelos TRI

En el dominio de la medición psicológica, el significado más adecuado de modelo es una combinación numérica de variables para predecir una variable dependiente. En la teoría clásica de los tests las variables independientes son la puntuación verdadera y el error de medición, y la variable dependiente la puntuación total (observada) en un test. Las variables independientes se combinan directa y aditivamente para predecir la variable dependiente, como se desprende de la ecuación básica de la TCT (PO = PV + E), revisada en el apartado precedente (Embretson y Reise, 2000).

Los modelos TRI pueden describirse con diferentes tipos de funciones logísticas que relacionan el nivel del atributo que se está midiendo con la probabilidad de responder de una manera determinada a los ítems. Estas funciones, representadas gráficamente, se denominan “función característica del ítem” o “cur-

va característica del ítem" (CCI). La curva característica de cualquier ítem indica la probabilidad que tienen las personas que se enfrentan a él de acertarlo (es decir, responder correctamente) (Hambleton y Rogers, 1991). Esta curva es un modelo teórico y debe verificarse mediante análisis apropiados (prueba de chi cuadrado, por ejemplo) que se ajusta a los datos reales (observados).

El análisis de un ejemplo propuesto por Abad y colaboradores (2006) nos permitirá una mejor comprensión de las afirmaciones precedentes. Imaginemos un test de inteligencia que ha sido aplicado a una muestra y cuya puntuación menor y mayor es 50 y 150, respectivamente. A continuación identificamos a las personas que han obtenido la puntuación 50 (supongamos que son 132). Por otro lado, estimamos cuántas personas con esa puntuación han acertado un ítem determinado (supongamos que 5 personas) y calculamos la proporción (5/132 = 0,04). Hacemos lo mismo con los que obtuvieron en el test 51 puntos (y obtenemos la proporción de aciertos a ese mismo ítem, supongamos que 0,15), 100 puntos (la proporción es 0,45), y 150 puntos (la proporción de aciertos a ese ítem es de 0,99). De este modo, podemos inferir que a mayor nivel en el rasgo (inteligencia, en el ejemplo), mayor es la proporción de aciertos en el ítem en cuestión.

En la aplicación de la TRI, un paso insoslayable (después de verificar que se cumplan los supuestos de unidimensionalidad e independencia local) es optar por un modelo teórico que suministre una buena representación del rendimiento en los ítems. Los modelos más utilizados en la práctica son: el modelo de Rasch o de un parámetro (*b*, dificultad), el modelo de dos parámetros (*b* y *a*, dificultad y discriminación), y el modelo de tres parámetros (dificultad, discriminación y adivinación, o *b*, *a* y *c*). Estos modelos tienen en común el uso de la CCI para especificar la relación entre el rendimiento observado en los ítems de un test y los rasgos o aptitudes latentes que explican ese desempeño.

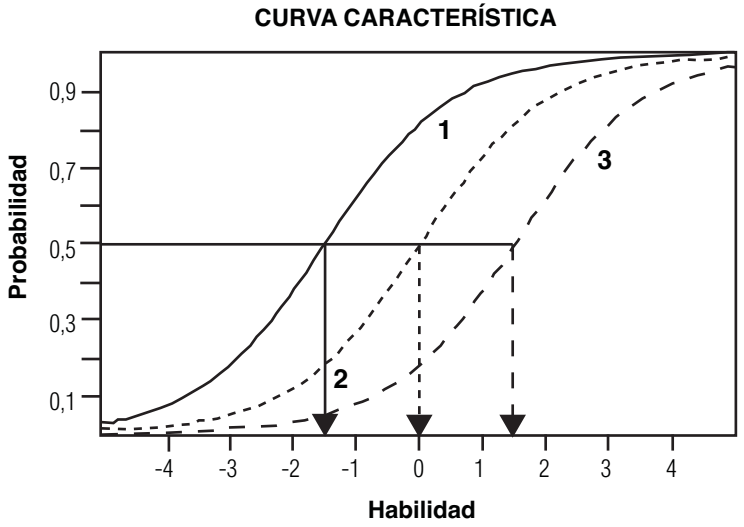
El más sencillo es el de un parámetro (1P) o modelo de Rasch (1963), que describiremos más detalladamente. En este modelo la probabilidad de acertar un ítem depende solamente del nivel de dificultad de dicho ítem y del nivel del individuo en la variable medida (nivel de aptitud o rasgo). Por consiguiente, el modelo de un parámetro nos indica qué nivel de aptitud requiere un ítem para responderlo correctamente (Martínez Arias, 1995).

La expresión matemática es la siguiente:

$$P\ (\theta) = \frac{e^{D(\theta - b_j)}}{1 + e^{D(\theta - b_j)}}$$

En la fórmula precedente, $P\ (\theta)$ es la probabilidad de acertar un ítem cuando la persona tiene un determinado nivel de rasgo o habilidad (θ), que normalmente asume valores entre –3 y +3. El símbolo *b* se refiere al parámetro de dificultad del ítem, que en la práctica suele expresarse en una escala con media 0, desviación estándar 1 y rango de valores entre -3 y 3, puesto que se mide en la misma escala que θ. Los símbolos *D* y *e* son valores constantes. En la figura siguiente podemos examinar la curva característica de tres items.

Figura 9.1. Curva característica de tres ítems. Modelo 1P



En el eje de abscisas se representan los diferentes valores del rasgo que mide el ítem (habilidad) y en el eje de ordenadas la probabilidad de acertar el ítem. Los tres ítems del gráfico presentan diferentes niveles de dificultad y de habilidad (theta), cuando la probabilidad de responder correctamente al ítem es igual a 0,5.

El parámetro de dificultad (*b*) es el puntaje en la escala del rasgo (θ) cuya probabilidad de respuesta correcta es igual a 0,5. De lo anterior se deriva que si el nivel del rasgo en un individuo es superior al requerido por el ítem será mayor la probabilidad de responder correctamente; y si el nivel de rasgo (o aptitud) del individuo es inferior al requerido por el ítem (su nivel de dificultad) será mayor la probabilidad de que responda incorrectamente.

En el primer ítem a la izquierda del gráfico, el valor θ al que corresponde la probabilidad 0,5 de acertar es, aproximadamente, igual a –1,50; por lo tanto, la dificultad del primer ítem (b) es igual a –1,50 (la persona que tiene un theta de –1,50 tiene una probabilidad del 50% de acertar ese ítem). En el segundo ítem, el valor de θ al que corresponde P (θ) = 0,5 es aproximadamente 0,00, y por lo tanto, *b* = 0 (el individuo que se encuentra en el punto medio de la escala del rasgo posee una probabilidad del 50% de acertar ese ítem). Finalmente, el ítem 3 es el más difícil puesto que su valor *b* es de 1,50. La gráfica muestra que la probabilidad de acertar el ítem decrece sistemáticamente desde el ítem 3 hasta el ítem 1 (el más fácil). También podemos observar que al aumentar la dificultad de un reactivo su curva se corre hacia la derecha. Por lo tanto, de la figura anterior podemos inferir tres propiedades importantes de este modelo:

a) Cuando una persona responde a un ítem en su nivel de competencia, tendrá la misma probabilidad de emitir una respuesta correcta que una incorrecta. Dicho de otro modo, la dificultad de un ítem es el valor r cuando P (θ) = 0,5.
b) Si el nivel θ es extremadamente bajo, la probabilidad de acierto se aproxima a 0. Es decir, este modelo considera que no se producen aciertos por azar.
c) La pendiente (inclinación) que tiene la curva del ítem en el parámetro *b* es la misma para cualquier ítem y se relaciona con la discriminación del ítem. Este modelo de un parámetro considera que todos los ítems tienen la misma capacidad discriminativa.
d) Según los propósitos de la medición, uno podría seleccionar ítems más dificultosos o fáciles.

Se ha señalado que el modelo de Rasch es el único que conduce a una auténtica medición en psicología (véase cap. 1) y, por consiguiente, es el más recomendable cuando los supuestos de igual discriminación y no adivinación de las respuestas pueden verificarse (Embretson y Reise, 2000). Sin embargo, el supuesto de igual capacidad discriminativa de los ítems algunas veces es difícil de corroborar, excepto en tests que miden áreas muy específicas de conocimiento o aptitud. Por otro lado, también asume que no hay respuestas correctas por adivinación, algo problemático de comprobar cuando los tests son utilizados en dominios aplicados en los que los resultados son importantes para la vida real de los examinados (selección de personal, por ejemplo). Por consiguiente, se han desarrollado dos modelos adicionales para enfrentar esas dificultades del modelo de un parámetro.

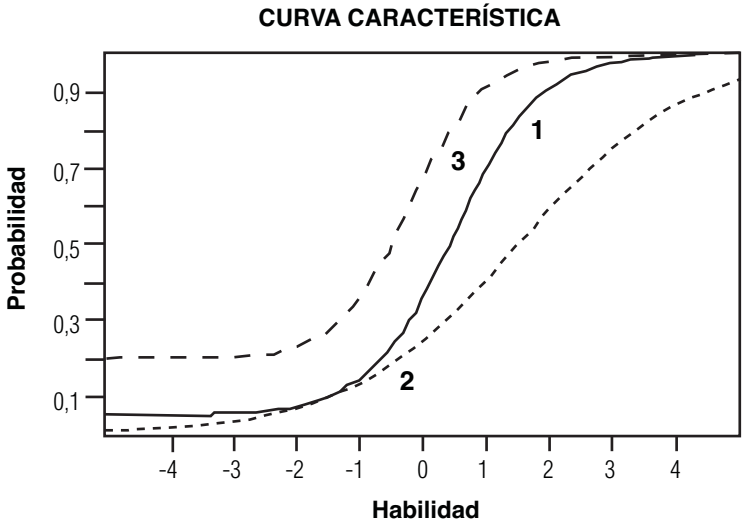
El modelo logístico de dos parámetros (2P) añade al anterior un segundo parámetro relacionado con la capacidad discriminativa del ítem. Esta propiedad indica hasta qué punto un ítem puede diferenciar entre los examinados que poseen habilidades bajas y altas, en un nivel de dificultad (parámetro *b*) determinado del ítem. La capacidad discriminativa de un ítem se simboliza con *a* y se refleja en la inclinación o pendiente de la curva del ítem. Normalmente estos valores varían entre 0,3 y 2,5, y se consideran ítems muy discriminantes aquellos que poseen valores superiores a 1,34, moderadamente discriminante entre 0,65 y 1,33, y escasamente discriminantes los valores de 0,64 o inferiores. Los ítems cuya curva posea mayor pendiente (se incline con mayor rapidez) son más útiles para separar (discriminar) los diferentes niveles de rasgo de la muestra que los que tienen una curva de menor pendiente (inclinación más suave, menos abrupta).

Por último, el modelo logístico de tres parámetros (3P) añade a los anteriores (*a* y *b*) un tercer parámetro, *c*, que representa la probabilidad de acertar un ítem por adivinación en los examinados con bajo nivel de aptitud. Si, por ejemplo, un ítem posee cin-

co alternativas de respuesta y una es la correcta, existe una probabilidad del 20% de responder correctamente únicamente por adivinación. Los valores de *c* varían de 0 a 1, y valores superiores a 0,35 no son considerados aceptables. El valor de *c* indica la probabilidad de acertar el ítem por azar en todos los niveles de rasgo de la escala; así, por ejemplo, un c = 0,12 indica que existe un 12% de probabilidad de responder correctamente al ítem por adivinación o azar. En la figura 9.1, el ítem 1 tiene un valor *c* ligeramente superior que los otros dos (es relativamente más probable acertarlo por azar) puesto que la asíntota inferior (punto de origen inferior de la curva) de su curva característica es mayor a 0 (aproximadamente 0,25).

El gráfico siguiente presenta la curva característica de ítems, evidenciando sus parámetros *a*, *b* y *c*. El ítem 1 presenta valores de *a* = 1,50; *b* = 0,5 y *c* = 0,05; el ítem 2 valores de *a* = 0,75; *b* = 1,5 y *c* = 0,00; y el ítem 3 valores de *a* = 1,75; *b* = 0,20 y *c* = 0,20.

Figura 9.2. Curva característica de tres ítems. Modelo de tres parámetros (3 P)



Como dijimos antes, además de estos tres modelos tradicionales (y apropiados para tests de aptitud con ítems dicotómicos) existen otros que pueden ser utilizados en la evaluación de la personalidad y constructos relacionados, en los que se utiliza preferentemente un formato *likert* de respuesta. El más popular es el modelo de respuesta graduada (Samejima, 1973), que es una extensión del modelo 2 P. También existen programas estadísticos para obtener los parámetros de los ítems y el rasgo latente en estos modelos. En el texto de Susan Embretson y Steven Reise *(Item response theory for psychologists*, 2000) se presenta una exposición muy comprensiva de estos modelos para ítems politómicos.

La ventaja esencial de los parámetros de las propiedades psicométricas de los ítems (dificultad, discriminación y adivinación), estimados a partir de las curvas características de los ítems, es que son más estables que los estimados con los métodos de la teoría clásica de los tests (la proporción de respuestas correctas a un ítem o la correlación ítem-test, por ejemplo) y menos dependientes de las muestras y los instrumentos utilizados.

9.4. Estimación de parámetros

Una vez seleccionado un modelo de TRI, hay que administrar el test a una muestra grande y estimar los parámetros de los ítems (*a, b, c*) requeridos según el modelo utilizado (1P, 2P o 3P) y los parámetros del rasgo latente o theta (θ) de los sujetos, a partir de la matriz de respuestas obtenidas.

Si tenemos, por ejemplo, diez ítems que miden un mismo rasgo, los podemos aplicar a una muestra de 300 personas. La matriz de datos tendrá 300 filas, siendo cada fila la secuencia de 1 (aciertos) y 0 (errores) de cada persona de la muestra. Si queremos aplicar el modelo logístico de tres parámetros, tendremos que estimar los 30 parámetros de los ítems (es decir, *a*, *b* y *c* de cada ítem) y 300 parámetros de las personas (los 300 valores de θ, uno por persona).

La estimación de parámetros es el paso que nos permite estimar los valores desconocidos de los parámetros de los ítems y de los niveles de rasgo, partiendo de las respuestas a los ítems observadas en las personas.

Para obtener estas estimaciones de parámetros se aplica el método de máxima probabilidad, con algunas variantes que no analizaremos aquí. La lógica general de la estimación consiste en encontrar los valores de los parámetros que hagan más probable (verosímil) la matriz de respuestas obtenida. Por ejemplo, si lanzamos una moneda diez veces y obtenemos siete caras, el estimador más probable del parámetro P (probabilidad de sacar cara al tirar la moneda) es 7/10 = 0,7. Este estimador proporciona el valor de P bajo el que tiene máxima probabilidad el suceso que hemos encontrado. En TRI, el procedimiento de estimación sigue una lógica similar y el estimador se simboliza con la letra L. Se obtienen las estimaciones de los parámetros de los ítems y de los niveles de θ con los que la matriz de datos observada (proporciones de aciertos y errores) tiene la máxima compatibilidad.

A continuación se ejemplifica cómo estimar el nivel de rasgo de un individuo en un test utilizando el estimador L de máxima probabilidad. Supongamos que una persona ha realizado un test de aptitud matemática que incluye tres ítems, emitiendo dos respuestas correctas y una incorrecta. Según el modelo de Rasch sabemos que los parámetros de dificultad (*b*) de esos ítems son -1, 0 y 2. Aplicando la regla de multiplicación de probabilidades, la fórmula para obtener L (estimador de máxima probabilidad) es:

$$L = (P1)\,(P2)\,(Q)$$

Donde P1 es la probabilidad de acertar el ítem 1, P2 la probabilidad de acertar el ítem 2 y Q (1-P) la probabilidad de errar el ítem 3.

En la tabla siguiente tenemos los ítems del test y sus parámetros de dificultad, en las dos primeras columnas. Las columnas que siguen representan la probabilidad de acertar cada uno de los ítems según determinados niveles de theta y el nivel de dificultad de los ítems.

Tabla 9.2. Estimación del nivel de rasgo. Método de máxima probabilidad

		Theta						
Ítem	**b**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**
1	-1	0,02	0,15	0,50	0,85	0,96	0,99	0,99
2	0	0,01	0,02	0,16	0,49	0,85	0,96	0,98
3	1	0,01	0,01	0,15	0,15	0,49	0,85	0,97

Aplicando la fórmula de L al ejemplo, tenemos que:

$$L\,(-3) = 0{,}02 \,.\, 0{,}01 \,.\, (1-0{,}01) = 0{,}00$$

$$L\,(-2) = 0{,}15 \,.\, 0{,}01 \,.\, (1-0{,}01) = 0{,}00$$

$$L\,(-1) = 0{,}50 \,.\, 0{,}16 \,.\, (1-0{,}15) = 0{,}06$$

$$L\,(0) = 0{,}85 \,.\, 0{,}49 \,.\, (1-0{,}15) = 0{,}35$$

$$L\,(1) = 0{,}96 \,.\, 0{,}85 \,.\, (1-0{,}49) = 0{,}41$$

$$L\,(2) = 0{,}99 \,.\, 0{,}96 \,.\, (1-0{,}85) = 0{,}14$$

$$L\,(3) = 0{,}99 \,.\, 0{,}98 \,.\, (1-0{,}97) = 0{,}02$$

De este modo, de los siete valores de theta (nivel del rasgo) considerados, el nivel de aptitud matemática estimado para ese individuo es igual a 1, por ser el L (el estimador de probabilidad) correspondiente más elevado (0,41). Expresado de otra manera, 1 es el puntaje total (nivel del rasgo) más probable de este individuo en este hipotético test de matemática.

En el ejemplo anterior se ha estimado la aptitud conociendo los parámetros de los ítems; para ello se ha administrado el test a una muestra y se ha obtenido el estimador L de theta a partir del patrón de respuesta a los ítems del test. Recíprocamente, también podríamos haber estimado los parámetros de los ítems conociendo solamente el nivel de aptitud de los examinados. Por

ser más complejos los cálculos requeridos no se ejemplificará este último procedimiento, pero la lógica es la misma.

Es fácilmente imaginable que, cuando se aplican los modelos TRI a muestras reales, estas estimaciones de parámetros se realizan mediante programas estadísticos de computación. El procedimiento de estimación de parámetros sigue la siguiente secuencia: a) se aplica el test a una muestra grande, b) se obtiene la matriz de respuesta (aciertos y errores) correspondiente a cada ítem del test, y c) mediante un software estadístico adecuado se estiman los parámetros de los ítems y los niveles del rasgo de la muestra (Abad y colaboradores, 2006).

Para que los parámetros sean estables, es necesario utilizar muestras grandes (n > 500), las que posibilitan el ajuste a cualquier modelo de los mencionados anteriormente. Cuando se trabaja con muestras más pequeñas el mejor modelo es el de Rasch, y por ese motivo es uno de los más populares

9.5. Función de información del ítem y del test

Una vez administrado el conjunto de ítems de un test y obtenido sus parámetros así como el nivel de habilidad de cada persona de la muestra, la TRI nos permite estimar la precisión y el error típico de medición para diferentes niveles del rasgo. La función de información, en TRI, es el indicador de la precisión de una medición en un punto de la escala de medición del rasgo latente (0, 1, -1, etc.), a diferencia de la TCT que asume que la confiabilidad (y el error estándar de medición) es igual para todas las personas evaluadas por un test (Abad y colaboradores, 2006).

La función de información del ítem (FI) depende esencialmente de la pendiente de la curva (capacidad de discriminación del ítem) y del error estándar de medición (cuanto más pequeño, mayor información). Es decir, cuanto mayor sea la pendiente y menor el error de medición de un ítem, mayor será la información. Por lo general, los ítems dan su mayor nivel de información en los valores del rasgo latente próximos a su nivel de dificultad. Cuando la pendiente de la CCI es grande, cambios mínimos en la aptitud se reflejarán en cambios considerables en la probabilidad de acertar el ítem.

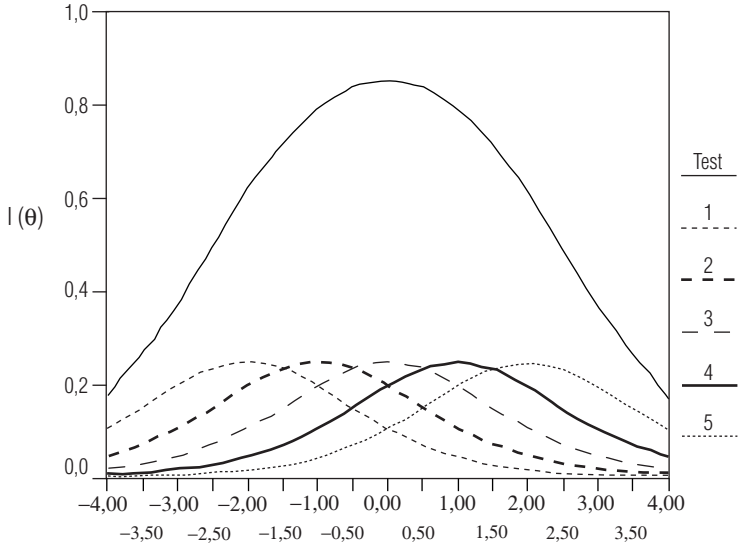
Como dijimos en el capítulo 3 sobre confiabilidad, otra característica importante de un ítem es la magnitud del monto de error de medición o la variabilidad de sus puntuaciones en un determinado nivel de theta (Martínez Arias, 1995). El error de medición estándar, entonces, en el marco de la TRI, no es uniforme para todos los niveles de aptitud (como en la TCT) y se expresa por la desviación estándar del ítem en un nivel de rasgo puntual. En consecuencia, a menor varianza de un ítem en determinado nivel de aptitud, mayor será la información aportada y la precisión de la medición.

Debe considerarse que la función de información del test en su totalidad (FIT) proviene de la sumatoria de la función de información de cada uno de sus ítems. Cuanto mayor sea el valor FIT menor será el valor del error típico de estimación y, por consiguiente, mayor la precisión de la medición. La función de información de un test en su conjunto, y por lo tanto el error típico de medida, para un determinado nivel de rasgo, depende fundamentalmente de: a) los parámetros de discriminación de los ítems (cuanto mayores sean los parámetros a, mayor será el valor de la información); b) los parámetros de adivinación (cuanto más bajos sean los valores de c, mayor será la información); c) la cantidad de ítems que tenga (suponiendo que los ítems tengan las mismas propiedades psicométricas, a mayor longitud mayor información) y d) la convergencia entre el nivel de rasgo θ y los parámetros de dificultad (b) de los ítems (cuanto más próximos sean, mayor será el monto de información aportado). En general, los valores inferiores a 1 de la FIT son indicadores de baja información (precisión), entre 1 y 1,69 son valores moderados y se consideran valores elevados los de 1,70 o superiores. La función de información de un test es mucho más alta que la FI de los ítems específicos de ese test. Un test siempre mide de manera más precisa que un ítem (Martínez Arias, 1995).

En la figura siguiente se representan las funciones de información de 5 ítems y de un test, utilizando el modelo 1P (de dificultad o modelo Rasch). En el eje vertical (ordenada) se representan los valores de la FI (del test e ítems), y en el eje horizontal (abscisa) el nivel de rasgo. A partir de la gráfica podemos inferir que el test resulta más informativo (confiable) en los valores medios del rasgo que en los niveles extremos, algo

bastante común en la mayoría de los tests. El pico de la función de información en algún punto de la escala de aptitud indica que el test mide con mayor precisión en los niveles próximos a ese pico. El gráfico también muestra claramente que el test es más preciso (su función de información es más elevada) que cualquiera de sus ítems.

Figura 9.3. Función de información de los ítems y el test



En síntesis, la precisión de una medición realizada mediante tests, en el contexto de la TCT, se verifica mediante un índice global de confiabilidad y de error estándar de medición. Por el contrario, en el contexto de la TRI, la precisión es evaluada por un índice condicional denominado "función de información del test". El término "condicional" refiere a que puede variar para diferentes valores del rasgo latente, al igual que el error típico de medida.

9.6. Funcionamiento diferencial del ítem

Algunos tests son criticados por estar sesgados respecto a las minorías étnicas o de otro tipo, y una de las ventajas de la teoría de la respuesta al ítem es que proporciona un marco de referencia unificado para interpretar los sesgos en el nivel de los ítems, obteniendo una estimación del funcionamiento diferencial del ítem (FDI). Esta última es una de las más importantes aplicaciones de la TRI, junto a los tests adaptativos computarizados y los bancos de ítems para tests de rendimiento. En el capítulo precedente ya suministramos ejemplos de investigaciones que analizaban el FDI. Aquí revisaremos los fundamentos de esta metodología en el contexto de la teoría de respuesta al ítem.

Un ítem es sesgado cuando individuos con un mismo nivel de habilidad o rasgo latente (asertividad, por ejemplo), pero que pertenecen a distintos grupos sociales, culturales o étnicos no tienen la misma probabilidad de responderlo correctamente. Ya a comienzos del siglo pasado, Binet observó que los niños de nivel socioeconómico bajo se desempeñaban mal en algunos ítems, atribuyendo este hecho al entrenamiento cultural y no a la inteligencia *per se* (rasgo latente medido por su escala). El ítem "Lloro fácilmente" de un inventario de personalidad, por ejemplo, puede conducir a una medición sesgada del constructo depresión, desfavorable para las mujeres, puesto que los hombres (como población) están culturalmente condicionados a no expresar tanto sus emociones, al menos en la mayoría de las culturas. Si un test estuviera compuesto por varios ítems semejantes al anterior, las mujeres sistemáticamente serían diagnosticadas como más depresivas que los hombres aunque posean el mismo nivel real de ese rasgo.

La presencia de FDI indica que un ítem posee parámetros (dificultad, adivinación y/o discriminación) diferentes en grupos distintos, y puede evidenciar una violación del supuesto de equivalencia que deben cumplir los tests para no producir una medida culturalmente sesgada, tal como vimos en el cap. 7, "Adaptación de los tests a otras culturas". Por lo tanto, el FDI se identifica comparando los parámetros de los ítems en grupos diferentes. El FDI, estrictamente considerado, sólo indica que hay diferencias significativas en el rendimiento de grupos diferentes

en un ítem, pero no es un indicador categórico de sesgo. Por ejemplo, los varones superan a las mujeres (como población) en tests de razonamiento espacial y las mujeres son superiores en su rendimiento promedio en tests verbales. Estas diferencias, si bien podrían ser explicadas por diferencias biológicas y/o culturales entre los dos sexos, no indican necesariamente que los ítems de esos tests favorezcan sistemáticamente a uno de los grupos en particular. Cuando las diferencias de rendimiento en el puntaje total de un test por parte de grupos diferentes obedece a diferencias reales en el constructo medido, y no a un sesgo de medición, se habla de "impacto del test".

Las propiedades de los ítems en la TCT (dificultad, discriminación) pueden utilizarse para comparar la respuesta a los ítems por parte de grupos diferentes. No obstante, estos índices pueden variar en diferentes puntos del continuo de una escala. Cuando se utilizan los parámetros de la teoría de respuesta al ítem, la estimación del FDI es más precisa.

La hipótesis nula de que los parámetros de respuesta de un ítem son iguales se puede formalizar, en el modelo 3P, de la siguiente manera:

$$H0;\ b1 = b2;\ a1 = a2;\ c1 = c2$$

Donde 1 y 2 son los dos grupos que se comparan, así como a, b, y c los parámetros de discriminación, dificultad y adivinación, respectivamente. Para contrastar la hipótesis nula (H0) es necesario conocer los parámetros de los ítems, así como las matrices de variancia y covariancia. Esto permite estimar la matriz de información para cada grupo y realizar una prueba estadística de chi cuadrado ($\chi 2$), o semejante, con la finalidad de contrastar esa hipótesis.

Una de estas medidas de asociación utilizadas para verificar el FDI es la razón de los logaritmos de las probabilidades desarrollada por Mantel y Haenszel (1959) y aplicada al estudio del FDI por Holland y Thayer (1988). La lógica de esta prueba es comparar el desempeño de dos grupos en el ítem, uno focal (el que concentra el interés de la investigación, el minoritario, el que se supone afectado por el FDI) y otro de referencia (el grupo estándar de comparación), para verificar si la probabilidad de

responder correctamente al ítem es igual en ambos grupos o no (indicando la existencia de FDI). Técnicamente, la hipótesis nula expresa que no hay diferencias entre los dos grupos en su probabilidad de respuesta correcta al ítem.

Para verificarla, se divide a los grupos focal y de referencia en subgrupos de acuerdo a la puntuación total que han obtenido en el test, agrupada en intervalos. Para cada ítem del test se determina la frecuencia de individuos en ambos grupos que lo realizaron de manera correcta e incorrecta (Hogan, 2004).

Tabla 9.3. Datos para el análisis Mantel-Haenszel del FDI

Intervalo de puntaje total	1-10		11-20		21-30	
Rendimiento en el ítem 1	+	-	+	-	+	-
Grupo de referencia	15	12	20	20	10	5
Grupo focal	8	10	15	15	4	1

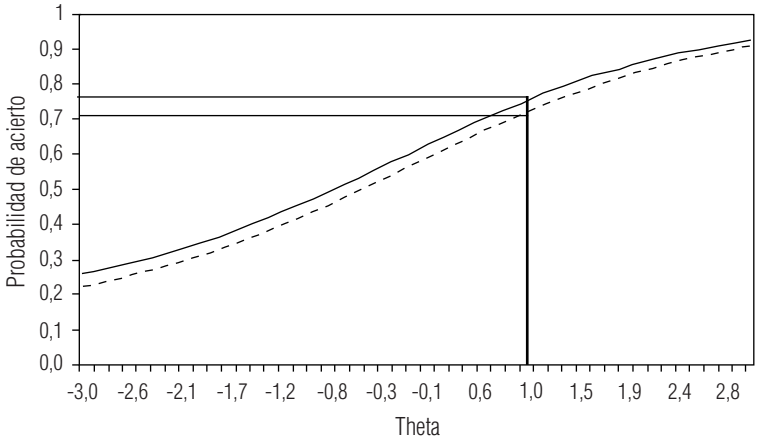
+ = Frecuencia de respuestas correctas

- = Frecuencia de respuestas incorrectas

En la tabla anterior se presentan datos de un test de 30 ítems dicotómicos, donde se analiza específicamente el FDI del ítem 1. Por ejemplo, en el intervalo de puntuación de 1 a 10 del test, 15 personas del grupo de referencia han respondido correctamente al ítem 1 y 12 incorrectamente. En el mismo intervalo, 8 personas del grupo focal emitieron una respuesta correcta y 10 no acertaron el ítem. Posteriormente, debe determinarse si la razón de las respuestas correctas en relación con las incorrectas es igual en los dos grupos, mediante el estadístico Mantel-Haenszel, basado en chi cuadrado. Los valores significativamente diferentes de 0 estarían indicando que uno de los grupos posee mayores posibilidades de éxito en un ítem y, por consiguiente, sugerirían la existencia de FDI. En la actualidad se utilizan también otros métodos, tales como el análisis factorial confirmatorio o la regresión logística, y *software* específico para analizar el FDI.

Otra forma de identificar la presencia de FDI es comparando directamente las curvas en lugar de sus parámetros. En este tipo de comparación, la evidencia de FDI se presenta cuando el área entre las dos CCI es diferente. En el gráfico siguiente puede apreciarse cómo un ítem de un test de rendimiento en matemática difiere en su parámetro de dificultad (*b*) en dos grupos diferentes (varones y mujeres, por ejemplo).

Figura 9.4. Representación de las CCI de un ítem con FDI



Se puede observar que, para una misma magnitud de θ, el valor P (θ) es siempre superior para los varones, es decir que niveles iguales de competencia en la variable medida (θ) no se corresponden con probabilidades iguales de responder exitosamente al ítem. En este caso, el ítem tiene FDI desfavorable a las mujeres (Línea Azul), pues los valores P (θ) para un mismo nivel de aptitud son siempre mayores para los varones (Línea Roja). Por ejemplo, para un nivel de aptitud θ = 1,0 la probabilidad de éxito en el ítem para las mujeres es de 0,71 mientras que para los varones es de 0,76.

APÉNDICE

ANÁLISIS PSICOMÉTRICOS CON SPSS

Marcos Cupani

Hasta hace aproximadamente tres décadas, la mayoría de los análisis estadísticos requeridos para estudios psicométricos (confiabilidad, validez, baremación) se hacían con el auxilio de calculadoras manuales. En la actualidad se dispone de una variedad de programas computarizados que simplifican enormemente estas tareas y permiten la realización de análisis cada vez más poderosos y sofisticados.

En este apartado nos referiremos a algunos análisis psicométricos que emplean el SPSS (Statistical Package for Social Sciences, 1995), un programa que se ejecuta desde una computadora personal con entorno Windows. Existen otros programas para realizar análisis más específicos, tales como el BILOG para la teoría de respuesta al ítem, y EQS, AMOS y LISREL para el modelo de ecuaciones estructurales (análisis factorial confirmatorio, por ejemplo). También se dispone de programas generales semejantes a SPSS y con funciones similares, tales como SAS o STATISTICAL.

La intención de este capítulo es introducir al lector en los fundamentos de los procedimientos analíticos, de modo de que cuando utilice un programa estadístico asistido por computadora sepa cómo usar la información que éste genera. Aunque no es necesario un conocimiento profundo de estadística para utilizar el SPSS, comprender los procedimientos estadísticos que subyacen a los diferentes análisis psicométricos facilita la interpretación adecuada de las salidas (*outputs*) del programa.

A continuación se revisarán algunas rutinas del SPSS incluyendo gráficos explicativos. En primer lugar se presenta el aná-

lisis correlacional (bivariado), aplicable a estudios de confiabilidad test-retest y formas equivalentes, así como a investigaciones que conducen a obtener evidencias de validez convergente-discriminante y de las relaciones entre un test y un criterio. Luego examinaremos una de las alternativas que nos ofrece SPSS para verificar la consistencia interna de un test. Por último, exploraremos uno de los métodos fundamentales en la psicometría contemporánea, el análisis de regresión múltiple, utilizado en estudios correlacionales con varios predictores y un criterio.

1. Correlación bivariada

1.1. Conceptos básicos

La mayoría de los métodos empleados para corroborar la confiabilidad y la validez de las puntuaciones de un test se interesan por el grado de consistencia o acuerdo entre dos conjuntos de puntuaciones obtenidas independientemente (Anastasi y Urbina, 1998). El coeficiente más utilizado es el de *correlación momento-producto de Pearson (r),* que expresa el grado de asociación lineal entre dos variables.

Para poder interpretar el coeficiente de correlación es necesario establecer una distinción inicial entre *estadístico* y *parámetro*. Un estadístico es una cifra que describe a una muestra, en cambio un parámetro es un valor que describe a una población. Por lo general, estimamos un valor correspondiente a la población a partir de la información obtenida en la muestra, es decir, tenemos un estadístico y queremos estimar un parámetro. Esta distinción entre estadístico y parámetro es importante para comprender las nociones de intervalo de confianza y de prueba de significación, conceptos muy utilizados para interpretar los estadísticos más usuales (Gardner, 2003).

Para un conjunto de datos cualesquiera, una vez calculado el coeficiente de correlación entre un par de variables, debe realizarse un sencillo test de hipótesis basado en la distribución t de Student. Esta prueba permite evaluar la significación del coeficiente de correlación y confirmar si existe o no una asociación

estadísticamente significativa (significativamente diferente de 0 y no producida por azar) entre ambas variables. Asimismo, puede obtenerse un intervalo de confianza para el coeficiente de correlación en la población.

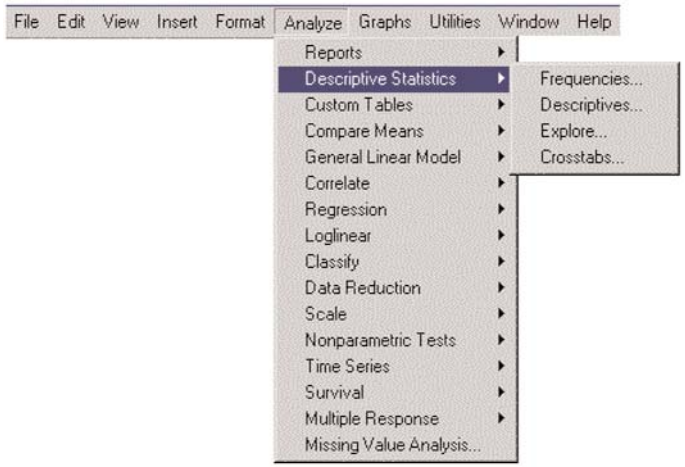
1.2. Un ejemplo de correlación bivariada aplicando el método test-retest

Si disponemos de las puntuaciones de una muestra de personas en un test y, después de transcurrido un tiempo, volvemos a medir a las mismas personas en el mismo test, cabe suponer que, si son confiables (estables) las puntuaciones de ese test, deberíamos obtener una correlación elevada entre ambos mediciones. La correlación entre la medición inicial o "test" y la evaluación posterior o "retest" se expresa mediante el *coeficiente de correlación,* el cual indicará mayor estabilidad de las puntuaciones cuanto más se aproxime a 1.

En una investigación se administró el Inventario de Autoeficacia para Inteligencias Múltiples (Pérez, 2001) en dos oportunidades a la misma muestra de 150 estudiantes de último año de la escuela media, con un intervalo de dos meses, y se calculó la correlación entre los puntajes obtenidos en la primera y segunda administración del test. Ilustramos este procedimiento con el análisis de la escala de Autoeficacia Lingüística de ese instrumento.

Como primer paso en cualquier análisis estadístico (sea bivariado o multivariado), es necesario formarse una idea lo más exacta posible acerca de las características de las variables en estudio y estimar las posibilidades de aplicar un estadístico especifico. Esto se consigue prestando atención a tres aspectos básicos: tendencia central (media, mediana y modo), dispersión (desviación típica, varianza, error típico de la media y amplitud) y forma de la distribución (asimetría y curtosis) de las variables. El usuario de SPSS puede acceder a estos análisis estadísticos, seleccionando la opción *Descriptivos*, del cuadro de diálogo *Analizar* que se puede observar en la figura siguiente.

Figura 1. Contenido del cuadro de diálogo "Analizar"



Los pasos son los siguientes:

1. Abrir la base de datos con la cual vamos a trabajar. En este caso es la *Base IAMI Test-Retest*, que contiene dos variables a ser analizadas: Test Lingüística (Ling1) y Retest Lingüística (Ling2).
2. Hacer clic en la opción *Analizar* y elegir la opción *Estadísticos Descriptivos* en la barra menú. De las nuevas opciones que se despliegan elegir *Descriptivos*
3. Se abrirá la ventana de *Descriptivos* con la lista de las variables del archivo de datos que poseen formato numérico. El usuario debe seleccionar las variables cuantitativas, en este caso las escalas del IAMI, y trasladarla a la lista de *Variables*.
4. Por defecto, este procedimiento permite calcular la media, desviación típica, valor mínimo y valor máximo. No obstante, al pulsar el botón *Opciones…*, y en el recuadro *Distribución*, podemos marcar las opciones asimetría y curtosis, y de esa manera obtener un panorama de la distribución de las variables que queremos analizar. El Visor de resultados suministra la información que recoge la tabla 1.

Tabla 1. Estadísticos de la opción *Descriptivos.*

Descriptive Statistcs

	N	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std.Error	Statistic	Std. Error
Test Lingüística	150	55,77	15,230	-0,239	0,198	-0,281	0,394
Retest Lingüística	150	51,38	14,748	-0,338	0,198	-0,204	0,394
Valid N (listwise)	150						

En la primera columna se puede observar el número de casos (N = 150). En las dos columnas siguientes se presentan, para cada unas de las variables, la Media (Mean) y la desviación típica (Std.). La media aritmética nos permite determinar, de todos los posibles valores numéricos de la escala, cuál es el valor promedio obtenido por esta muestra. La desviación típica es un estadístico que nos da una idea de la magnitud de las desviaciones de los valores respecto a la media.

A partir de la cuarta columna de la tabla se presentan los índices que expresan el grado de asimetría de la distribución. Los valores de asimetría pueden ser positivos o negativos. Si los índices son positivos, quiere decir que hay una cola larga que se extiende hacia la derecha (los valores más extremos se encuentran por encima de la media); en caso de que sean negativos, la cola se extiende hacia la izquierda (los valores más extremos se encuentran por debajo de la media). Los índices de asimetría próxima a cero indican simetría.

En la sexta columna se aprecian los índices de curtosis, que expresan el grado en que una distribución acumula casos en sus colas en comparación con los casos acumulados en las colas de una distribución normal con la misma varianza. Si la distribución es relativamente puntiaguda en la parte media y tiene colas relativamente altas (es decir, también algunos valores extremos), la curtosis será grande. Si la distribución tiene relativamente pocos valores en los extremos, la curtosis será pequeña. Así pues, las muestras con valores de curtosis próximo a 0 indican semejanza con la curva normal. Las que tienen valores positivos tienden a tener colas más altas, mientras que las que

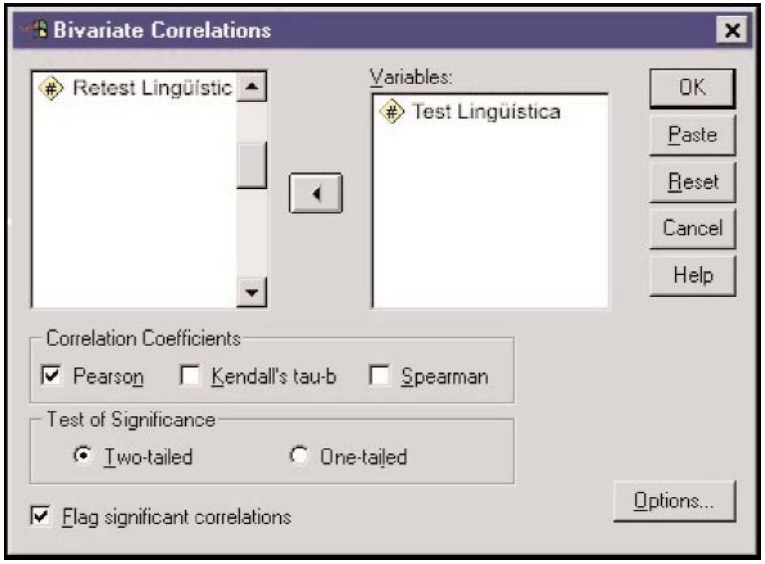
poseen valores negativos tienen colas más bajas. El programa también estima el error típico del índice de asimetría y de curtosis, respectivamente.

Una vez obtenidos el índice de asimetría y curtosis con sus respectivos errores típicos, podríamos usar esta información para probar si es razonable suponer que la muestra se extrajo de una población con distribución simétrica. Para ello, formaríamos una desviación normal estandarizada obteniendo el cociente del estadístico menos el parámetro, dividido por el error estándar. Por lo tanto, tomando los valores de simetría obtenidos en la variable Test Lingüística sería: –0,239– 0/0,198 = -1,209, se puede observar que el índice obtenido no es mayor que 1,96 (el valor que Z debe tener para ser significativo en un nivel 0,05), así que no tenemos pruebas que nos permita concluir que nuestra población esta sesgada. De la misma manera, podemos determinar si nuestra muestra tiene mayor o menor curtosis que una población normal, y entonces -0,281–0/0,394 = -0,713, donde el valor no es mayor a 1,96 y, por lo tanto, podemos concluir que no hay pruebas para suponer que la distribución de nuestra población es diferente de la normal. No obstante, los valores de asimetría y curtosis son negativos, situación que nos indica que la cola de la distribución se extiende levemente hacia la izquierda.

El coeficiente de correlación de Pearson puede ser calculado para cualquier conjunto de datos, pero el test de hipótesis sobre la correlación entre las variables requiere que al menos una de ellas tenga una distribución normal. Por consiguiente, examinadas las estadísticas descriptivas de las variables, y constatando que todas las variables presentan una distribución normal estamos en condiciones de analizar la correlación entre las dos variables.

El programa SPSS nos permite calcular diversos coeficientes de asociación, tales como el coeficiente de correlación Pearson, el Rho de Spearman y la Tau-b de Kendall (véase figura 2). Para las variables cuantitativas, normalmente distribuidas, es recomendable el coeficiente de correlación de Pearson. En casos en que los datos no estuviesen normalmente distribuidos, podríamos estimar los coeficientes Tau-b de Kendall o rho Spearman, que miden la asociación entre órdenes de rangos.

Figura 2. Cuadro de diálogo “Correlación Bivariada”



Para obtener el coeficiente de correlación, el usuario debe realizar los pasos siguientes:

1. Hacer clic en la opción *Analizar* y elegir la opción *Correlación* en la barra menú. Se abrirá otro menú a la derecha con las siguientes opciones: Bivariada…, Parcial… y Distancia…De las nuevas opciones que se despliegan, hacer clic en la opción *Bivariada*.
2. Se abrirá la ventana de *Bivariada…* con la lista de las variables del archivo de datos que poseen formato numérico. El usuario debe seleccionar las variables cuantitativas, y trasladarla a la lista de *Variables*.
3. Puede verse en la figura 2 que por defecto están seleccionadas las opciones *Pearson, bilateral* (*s*e podría haber seleccionado *unilateral* si se conoce de antemano la dirección de la asociación), y marcar las correlaciones significativas (el programa identifica las correlaciones significativas al nivel 0,05 con un solo asterisco y al nivel 0,01 con dos asteriscos).

4. En el Cuadro de diálogo *Opciones* podemos obtener información adicional: Medias, desviaciones típicas, los productos cruzados diferenciales y covarianzas.
5. Aceptando estas elecciones, el Visor de resultados ofrece la información que recoge la siguiente tabla.

Tabla 2. Coeficiente de estabilidad test-retest

Correlaciones			
		Test Lingüística	**Retest Lingüística**
Test Lingüística	Pearson Correlation	1	0,853 **
	Sig. (2-tailed)		0,000
	N	150	150
Retest Lingüística	Pearson Correlation	0,853**	1
	Sig. (2-tailed)	0,000	
	N	150	

** Correlation is significant at the 0,01 level (2-tailed).

La correlación entre las variables Test Lingüística y Retest Lingüística es positiva y significativa (0,853; *Sig. = 0,000*), lo que sugiere que los individuos que presentaron valores altos en las escalas del IAMI en la primera administración (Test) están asociados a valores altos de la segunda aplicación del test (Retest) y, a su vez, que los valores bajos en la primera administración están asociados a valores bajos en la segunda aplicación. En este caso se verifica una elevada estabilidad de las puntuaciones de la escala "Autoeficacia Lingüística".

Si bien el análisis de correlación bivariada aplicando SPSS se ha ilustrado con un ejemplo relacionado con la estabilidad temporal (test-retest) de las puntuaciones de un test, podría extenderse a todos los estudios psicométricos donde se correlacionen dos series de puntuaciones (formas equivalentes, relaciones test-criterio, correlación ítem-test, evidencia de convergencia y

discriminación, partición de mitades), tal como se ha desarrollado en los capítulos correspondientes de confiabilidad y validez.

Por ejemplo, en la tabla siguiente se puede observar un estudio de evidencia de validez test-criterio entre la escala Lingüística del IAMI y el promedio final en la asignatura Lengua. Se ve que la correlación es positiva, modesta y significativa (0,277; Sig. = 0,001). Esto sugiere que, en términos probabilísticos, los adolescentes que se sienten seguros de realizar adecuadamente actividades relacionadas con la inteligencia lingüística poseen un rendimiento relativamente más elevado en Lengua.

Tabla 3. Correlación de Pearson entre la escala Lingüística y el promedio en Lengua (N = 139 adolescentes de la ciudad de Córdoba)

Correlations			
		Autoeficiencia Lingüística	**Promedio Lingüística**
Autoeficacia Lingüística	Pearson Correlation	1	0,277 **
	Sig. (2-tailed)		0,001
	N	139	139
Promedio Lenguaje	Pearson Coprrelation	0,277 **	1
	Sig. (2-tailed)	0,001	
	N	139	139

** Correlation is significant at the 0,01 level (2-tailed)

2. Coeficiente alfa de Cronbach

2.1. Conceptos básicos

En el apartado anterior aprendimos cómo estimar la estabilidad temporal de las puntuaciones de un test. Si queremos conocer el grado en que los distintos ítems de un test miden la misma variable, estaríamos verificando la consistencia interna de

las puntuaciones de ese test. El estadístico más popular para evaluar esta dimensión de confiabilidad es el coeficiente alfa de Cronbach. Este coeficiente refleja el grado de co-variación de los ítems (Muñiz, 2001): si los ítems covarían fuertemente, asumirá un valor cercano a 1, y si los ítems son linealmente independientes, asumirá valores cercanos a 0.

Recordemos que el IAMI en su versión original (Pérez, 2001) incluye 8 escalas obtenidas por análisis factorial y 64 ítems (8 ítems por escala). El usuario de la prueba debe responder utilizando un formato de 10 alternativas, desde 1 *No puedo realizar esa actividad* a 10 *Completamente seguro de poder realizar exitosamente esa actividad*. Para realizar este estudio de consistencia interna, el IAMI fue administrado a 525 estudiantes de último año de la escuela media (nivel Polimodal) de la ciudad de Córdoba, Argentina.

El programa SPSS nos ofrece distintos modelos para estimar la consistencia interna. Los más utilizados son la partición en mitades, en la que el programa divide la escala en dos partes y correlaciona dichas partes, y el Alfa de Cronbach, que se basa en la covariación de los ítems. También da la posibilidad de realizar análisis adicionales, tales como el coeficiente de correlación intraclase, el test de Hotelling y la prueba de aditividad de Tukey, entre otros.

Para poder estimar el coeficiente alfa, el usuario debe seleccionar la opción *Escala* del cuadro de dialogo *Analizar* que se puede observar en la figura 1 y hacer clic en *Análisis de confiabilidad.*

Los pasos son los siguientes:

1. Abrir la base de datos con la cual vamos a trabajar. En este caso es la *Base Confiabilidad y Factorial IAMI*, que contiene 64 variables (ítems del IAMI), de las cuales analizaremos los 8 Items de la escala Lingüística.
2. Hacer clic en la opción *Analizar* y elegir la opción *Escala* en la barra menú. De las nuevas opciones que se despliegan, hacer clic cen *Análisis de Confiabilidad.*
3. Se abrirá el cuadro de diálogo *Análisis de confiabilidad* con la lista de las variables del archivo de datos que poseen formato únicamente numérico. El usuario debe seleccionar las variables cuantitativas, en este caso los 8 ítems de la escala Linguística, y trasladarlas a la lista de *Variables*.
4. Por defecto, este procedimiento permite calcular el coeficiente alfa de Cronbach. No obstante, al acceder al cuadro de diálogo *Estadísticos* del Análisis de Confiabilidad (véase figura 4), y en el recuadro *Descriptivos* podemos marcar las opciones *Ítem, Escala, y Escala si se elimina el Ítem*. Los estadísticos incluyen la media de escala y la varianza si el ítem se ha eliminado de la escala, la correlación entre el ítem y la escala, y el coeficiente alfa si el ítem se ha eliminado de la escala. Además, en la opción *Inter-Ítem*, se pueden obtener las matrices de correlaciones o covarianzas entre los ítems. En este caso elegimos las opciones *Ítems, Escala si se elimina el ítem, y correlación Inter-ítem.*
5. Aceptando estas elecciones, el Visor de resultados ofrece la información que recogen las tablas siguientes.

La primera tabla que ofrece el programa es la que contiene el número de casos, que no fue incluida en este apartado. En la tabla siguiente se muestra el coeficiente alfa obtenido en la escala Lingüística, el mismo coeficiente con los ítems estandarizados y el número de ítems que fueron incluidos en el análisis.

Tabla 4. Coeficiente alfa de la escala Lingüística (IAMI)

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
0,845	0,845	8

Como puede inferirse de la información precedente, el valor de alfa es adecuado (superior a 0,80). En la tabla siguiente se presentan para cada uno de los ítems, la Media (Mean) y la desviación típica (Std.). Se ha comentado previamente que los

ítems deben tener la capacidad de poner de manifiesto las diferencias existentes entre los individuos ,y para ello, se deben conseguir ítems que maximicen la varianza del test, con una desviación típica superior a 1 y media situada alrededor del punto medio de la escala (Nunnally y Bernstein, 1995). En este caso se puede observar que el valor medio más alto corresponde al ítem 4 (6,94), el valor mínimo corresponde al ítem 2 (4,68), y la mayor variabilidad en las repuestas se observó en los ítems 5 (S = 2,649) y 1 (S = 2,623).

Tabla 5. Estadísticos descriptivos de los ítems del procedimiento confiabilidad

Item Statistics			
	Mean	**Std. Deviation**	**N**
ITEM1	4,86	2,623	525
ITEM2	4,68	2,561	525
ITEM3	5,55	2,572	525
ITEM4	6,94	2,241	525
ITEM5	5,95	2,649	525
ITEM6	5,28	2,375	525
ITEM7	5,79	2,394	525
ITEM8	6,51	1,987	525

A continuación se presentan las matrices de correlación entre los 8 ítems que componen la escala. Evidentemente, cuanto mayor es la magnitud del coeficiente de correlación entre los ítems más elevada será la confiabilidad de la escala. Adicionalmente podríamos haber obtenido el promedio de las correlaciones inter-ítems marcando la opción *correlación* en la opción resúmenes del cuadro de diálogo *Estadísticos*.

Se recomienda que en general el valor de la correlación media inter-ítem esté situado entre 0,15 y 0,50 (Briggs y Cheek, 1986). En este caso, la correlación media inter-íitems es de 0,405 (rango de 0,232 a 0,589).

Tabla 6. Matriz de correlación entre los ítems

Inter-Item Correlation Matrix								
	ITEM1	**ITEM2**	**ITEM3**	**ITEM4**	**ITEM5**	**ITEM6**	**ITEM7**	**ITEM8**
ITEM1	1,000	0,506	0,589	0,476	0,434	0,439	0,405	0,297
ITEM2	0,506	1,000	0,412	0,247	0,255	0,397	0,298	0,263
ITEM3	0,589	0,412	1,000	0,541	0,416	0,413	0,431	0,232
ITEM4	0,476	0,247	0,541	1,000	0,526	0,368	0,534	0,325
ITEM5	0,434	0,255	0,416	0,526	1,000	0,334	0,462	0,408
ITEM6	0,439	0,397	0,413	0,368	0,334	1,000	0,576	0,347
ITEM7	0,405	0,298	0,432	0,534	0,462	0,576	1,000	0,425
ITEM8	0,297	0,263	0,232	0,325	0,408	0,347	0,425	1,000

The covariance matrix is calculated and used in the analysis.

Finalmente, se presenta la media y la varianza si el ítem se ha eliminado de la escala, la correlación entre el ítem y la escala compuesta por los otros ítems, y el coeficiente alfa si el ítem se ha eliminado de la escala. En la primera columna se presenta la media de las puntuaciones totales de los ítems donde en la suma de estas puntuaciones eliminamos el ítem correspondiente, es decir que el valor 40,70 es la media de la variable sumando los ítems 2 al 8. La segunda columna incluye las variancias de esta variable "suma" así obtenida. La tercera columna presenta el coeficiente de correlación de Pearson entre cada ítem y el puntaje total de la escala. En la cuarta columna aparecen los cuadrados de los coeficientes de correlación múltiple entre cada ítem y el resto. Finalmente, en la quinta columna se presenta el coeficiente alfa que obtendríamos si eliminamos el ítem correspondiente. Sin embargo, la decisión de eliminar un ítem sólo porque aumenta la confiabilidad es problemática; se recomienda que para este tipo de decisiones los criterios teóricos sean al menos tan importantes como los resultados empíricos. En este caso el coeficiente alfa (0,845) no se incrementa con la eliminación de ningún ítem de la escala como puede apreciarse en la última columna de la tabla siguiente.

Tabla 7. Estadísticos descriptivos de los ítems

Item-Total Statistics					
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbachs's Alpha if item Deleted
ITEM 1	40,70	135,029	0,656	0,487	0,816
ITEM 2	40,88	145,709	0,480	0,316	0,840
ITEM 3	40,01	137,326	0,630	0,467	0,820
ITEM 4	38,62	143,461	0,623	0,469	0,822
ITEM 5	39,61	139,159	0,572	0,385	0,828
ITEM 6	40,28	143,010	0,586	0,414	0,826
ITEM 7	39,77	139,977	0,639	0,492	0,819
ITEM 8	39,05	155,344	0,457	0,261	0,840

3. Análisis de regresión múltiple

3.1. Conceptos básicos

El análisis de regresión múltiple es un método utilizado para analizar la relación entre una variable dependiente (criterio) y dos o más variables independientes (predictores). El procedimiento implica básicamente obtener la ecuación mínimo-cuadrática que mejor exprese la relación entre la variable dependiente y las variables independientes. Para poder aplicar este método, las variables utilizadas en el análisis deben ser métricas o apropiadamente transformadas, y debe definirse previamente cuál es la variable dependiente y cuáles son las independientes. Estos dos pasos deben ir acompañados de un registro previo del cumplimiento de los supuestos de Linealidad, Independencia, Homocedasticidad, Normalidad y No-colinealidad, que garantizan la validez del procedimiento (Tabachnick y Fidell, 2001). Existen varios métodos de análisis de regresión múltiple, pero los más empleados son el análisis de regresión jerárquica, el análisis de regresión por pasos o *stepwise* y el análisis de regresión estándar. A continuación se ejemplificará el método

de regresión jerárquica, que es el más empleado en la psicología contemporánea por otorgar mayor importancia a la teoría subyacente a un test determinado.

Un ejemplo de análisis de regresión jerárquica

En el análisis de regresión jerárquica las variables predictoras son ingresadas a la ecuación de predicción en el orden sugerido por la teoría de base del test. A continuación se revisa una investigación cuyo objetivo fue identificar la contribución realizada por la aptitud para la matemática y la autoeficacia para inteligencias múltiples (variables predictoras) para explicar la variabilidad del rendimiento académico en Matemática (variable criterio), en una muestra de estudiantes secundarios de la ciudad de Córdoba (Pérez, Cupani y Ayllón, 2005).

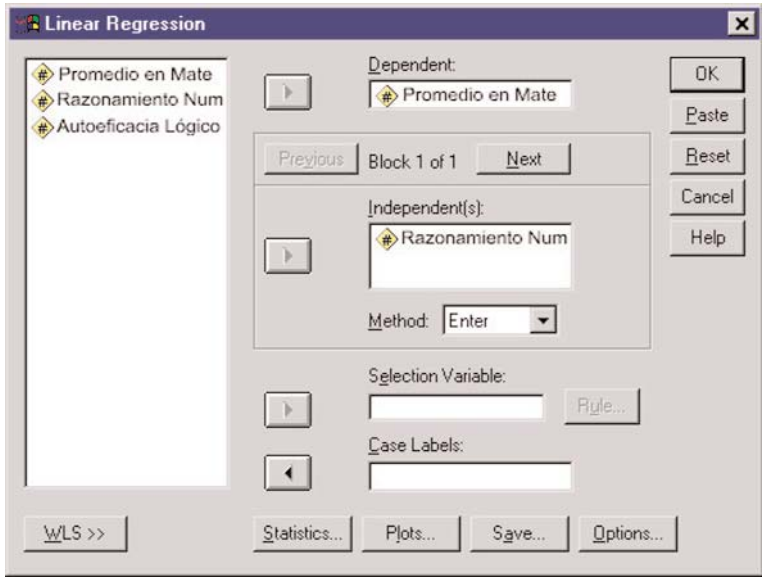
Con esa finalidad, se administró la escala de Autoeficacia Lógico-Matemática del IAMI y el subtest de Razonamiento Numérico del DAT-5 a una muestra de 138 estudiantes del último año de la escuela media (nivel Polimodal). Las variables independientes fueron incorporadas al modelo según los postulados de la teoría social cognitiva de carrera (Lent, Brown y Hackett, 1994): en primer lugar el test de aptitud cognitiva (uno de los antecedentes de la autoeficacia) y en segundo lugar la escala de autoeficacia. La variable criterio (rendimiento en Matemática) fue operacionalizada mediante el promedio anual en Matemática de los participantes.

Para realizar el análisis de regresión, el usuario debe seleccionar la opción *Regresión* del cuadro de diálogo *Analizar*. Luego, es necesario escoger la opción *Lineal…,* tal como se puede observar en el cuadro de diálogo de la figura 3.

Los principales estadísticos del análisis de regresión múltiple se encuentran en el cuadro *Estadísticos*, donde para cada modelo se pueden obtener: coeficientes de regresión, matriz de correlaciones, correlaciones parciales y semiparciales, *R* múltiple, *R* cuadrado, *R* cuadrado corregida, cambio en *R* cuadrado, error típico de la estimación, tabla de análisis de varianza, valores pronosticados y residuos.

Otro cuadro de diálogo que el usuario debe considerar es la opción *Métodos*, que permite especificar cómo se introducen las

Figura 3. Contenido del cuadro de diálogo *Analizar, Regresión*



variables independientes en el análisis. Los pasos para realizar el análisis de regresión jerárquica son los siguientes:

1. Abrir la base de datos con la cual vamos a trabajar. En este caso es la *Base Regresión*, que contiene 3 variables (Promedio en Matemática, Razonamiento Numérico y Autoeficacia Lógico-Matemática).
2. Hacer clic en la opción *Analizar* y elegir la opción *Regresión* en la barra menú. De las nuevas opciones que se despliegan marcar *Lineal*.
3. Se abrirá el cuadro de diálogo *Regresión Lineal* con la lista de las variables del archivo de datos que poseen formato numérico. El usuario debe seleccionar la variable dependiente, que en este caso es el Promedio en Matemática, y trasladarla a la lista de *Dependiente*.
4. En la opción método dejaremos la opción por defecto, que es *Introducir (enter)*, y como el procedimiento a utilizar es el análisis de regresión jerárquico, el usuario debe seleccionar la primera variable independiente que en este caso es Razonamiento Numérico, y trasladarla a la lista de *Independientes*. Para ingresar la segunda variable, el usuario debe hacer clic en la opción *Siguientes* y posteriormente trasladar la segunda variable, Autoeficacia Lógico-Matemática, a la lista de variables independientes.
5. Haciendo clic en el botón *Estadísticos*, se abrirá una nueva ventana, *Regresión lineal: Estadísticos*. Las opciones *Estimaciones* y *Ajuste del Modelo* estarán dadas por defecto, pero hay muchas otras opciones. En este caso se seleccionaron las opciones *Cambio del R cuadrado* y *Correlaciones Semiparcial y Parcial*.
6. Aceptando estas elecciones, el Visor de resultados ofrece la información que recogen las tablas 8, 9, 10 y 11.

La primera tabla que ofrece el programa es la que contiene el número de casos, que no fue incluida en este apartado. En la tabla siguiente se puede observar el coeficiente de correlación múltiple (R) y su cuadrado (R^2) o coeficiente de determinación. Este coeficiente expresa el porcentaje de varianza explicada de la variable dependiente por las variables independientes. En

Tabla 8. Resumen del modelo de regresión jerárquica

Model Summary

		Model 1	Model 2
R		0,446	0,496
R Square		0,199	0,246
Adjusted R Square		0,193	0,235
Std. Error of the Estimate		1,99046	1,93761
Change Statistics	R Square Change	0,199	0,047
	F Change	36,174	9,073
	df1	1	1
	df2	146	145
	Sig. F Change	0,000	0,003

nuestro ejemplo, el R^2 que debemos interpretar es el del último paso (Modelo 2), donde el valor de R^2 es de 0,246, que nos indica que el conjunto de variables predictoras explica aproximadamente un 25% de la varianza de rendimiento académico en Matemática.

En la misma tabla se observa, para cada uno de los pasos, el cambio experimentado por R^2 (R^2 cambio), con el ingreso a la ecuación de cada predictor, y el estadístico F con su respectiva significación (contrasta la hipótesis de que el cambio en R^2 vale 0). En nuestro ejemplo observamos que en el primer modelo, conformado por la variable Razonamiento Numérico del DAT, se obtuvo un R^2 de 0,199. En este primer paso no nos interesa analizar R^2 cambio, puesto que es igual al R^2. No obstante, sí podemos contrastar la hipótesis de que el valor poblacional de R^2 cambio es 0 mediante el estadístico razón de F. En este caso el valor de F es de 36,176, significativo al 0,000. En el segundo modelo, las variables predictoras son Razonamiento Numérico y Autoeficacia Lógico-Matemática, y el valor de R^2 aumenta hasta 0,246 (25% de la varianza explicada), produciendo la variable Autoeficacia Lógico-Matemática un incremento aproximado del 5% (R^2 cambio = 0,047; *F* = 9,076, *Sig.*= 0,003).

A continuación, en la tabla 9 se presentan los resultados del análisis de varianza, donde se describen tres fuentes de variación: Regresión, Residual y Total. Se puede observar la razón F que contrasta la hipótesis nula de que el valor poblacional de R

Tabla 9. Resumen del análisis de varianza

ANOVA						
Model		**Sum of Squares**	**df**	**Mean Square**	**F**	**Siq.**
1	Regression	143,320	1	143,320	36,174	0,000
	Residual	578,444	146	3,962		
	Total	721,764	147			
2	Regression	177,385	2	88,692	23,624	0,000
	Residual	544,380	145	3,754		
	Total	721,764	147			

es 0 y, por lo tanto, nos permite inferir si existe relación lineal significativa entre la variable dependiente y el conjunto de variables independientes tomadas juntas. En el modelo 2 se puede observar que el valor crítico (*Sig.* = 0,000) es significativo.

En la tabla siguiente se presentan los coeficientes de regresión de las variables incluidas en el modelo de regresión, es decir, la información necesaria para construir la ecuación de regresión para cada paso. Las primeras columnas recogen el valor de los coeficientes de regresión parcial (B) y su error típico. A continuación aparecen los coeficientes de regresión parcial estandarizados (Betas), los cuales proporcionan una estimación de la importancia relativa de cada variable dentro la ecuación de regresión. Las dos últimas columnas muestran el estadístico *t* y el nivel crítico (Sig.) obtenidos al contrastar las hipótesis de que los coeficientes de regresión parcial valen 0 en la población. Un nivel crítico por debajo de 0,05 indica que la variable contribuye significativamente a mejorar el modelo de regresión.

Tabla 10. Coeficientes de regresión

Coeficients						
		Unstandardized Coeficientes		**Standardized Coeficients**		
Model		**B**	**Std.Error**	**Beta**	**t**	**Siq.**
1	(Constant)	4,050	0,425		9.521	0,000
	Razonamiento Numérico	0,145	0,024	0,446	6,014	0,000
2	(Constant)	2,718	0,616		4,487	0,000
	Autoeficacia	0,112	0,026	0,345	4,342	0,000
	Lógico-matemática	0,033	0,011	0,239	3,012	0,003

En este caso, la ecuación de regresión para el pronóstico del rendimiento académico en Matemática sería igual a 2,718 + 0,112 + 0,033. Estos coeficientes no estandarizados se interpretan de la siguiente manera: el coeficiente correspondiente a la variable Razonamiento Numérico (0,112) indica que si el resto

de las variables se mantienen constantes, a un aumento de una unidad en Razonamiento Numérico le corresponde, en promedio, un aumento de 0,112 en rendimiento en Matemática. De modo análogo se interpreta el coeficiente no estandarizado (B) de la variable predictora autoeficacia lógico-matemática (0,033).

Los coeficientes Betas (coeficientes de regresión parcial estandarizados) son los que definen la ecuación de regresión cuando ésta se obtiene tras estandarizar las variables originales, es decir, luego de convertir las puntuaciones originales en puntuaciones típicas. En el análisis de regresión simple, los coeficientes Betas correspondientes a la única variable independiente coinciden exactamente con el coeficiente de correlación de Pearson. En cambio, en la regresión múltiple, los coeficientes de regresión estandarizados permiten valorar la importancia relativa de cada variable independiente dentro de la ecuación. Sin embargo, hay que señalar que estos coeficientes no son independientes entre sí; de hecho, se denominan "coeficiente de regresión parcial" porque el valor concreto estimado para cada coeficiente se ajusta teniendo en cuenta la presencia del resto de variables independientes.

A medida que se añaden más variables independientes a la ecuación, mayor consideración se deberá prestar a las intercorrelaciones entre las variables independientes. Si las variables independientes están correlacionadas, entonces comparten algo de su poder predictivo. Para poder estimar cuál es el efecto compartido, podemos calcular dos coeficientes adicionales de gran utilidad, la *correlación parcial (partial) y semiparcial (part)*. La primera (parcial) es la correlación entre una variable independiente (X_1) y una variable dependiente (Y) cuando se han suprimido (controlado) los efectos de la otra variable independiente (X_2) tanto en X_1 como en Y. La segunda (semiparcial) refleja la correlación entre la variable independiente (X_1) y una variable dependiente (Y) cuando se controlan los efectos de las variables independientes restantes del modelo sobre X_1. En el cuadro que sigue se presentan los coeficientes de correlación de orden 0, parcial y semiparcial de las variables independientes Autoeficacia Lógico-Matemática y Razonamiento Numérico con la dependiente Promedio en Matemática.

Tabla 11. Correlación parcial y semiparcial

Coeficients[a]

Model		Correlations		
		Zero-order	Partial	Part
1	Razonamiento numérico	0,446	0,446	0,446
2	Razonamiento numérico	0,446	0,339	0,313
	Autoeficacia Lógico-matemática	0,384	0,243	0,217

[a] Dependent Variable: Promedio Matemática.

Analizando los datos precedentes se puede constatar que existe una correlación significativa entre Autoeficacia Lógico-Matemática y promedio final en Matemática de 0,217, una vez que se ha excluido de Autoeficacia Lógico-Matemática cualquier variabilidad en común con los efectos combinados de Razonamiento Numérico. También podemos observar que esta correlación semiparcial es un poco más baja que la correlación original (bruta, de orden 0) entre Autoeficacia Lógico-Matemática y promedio en Matemática que era de 0,384. No obstante, en algunos casos la correlación semiparcial podría ser mayor que la de orden 0.

El análisis de regresión múltiple no puede discriminar adecuadamente los componentes de varianza explicada por una variable independientemente de las otras, ya que el R incremental tiende a subestimar el poder explicativo de los predictores ingresados después del primer predictor. Un método complementario que suministra una estimación más precisa de la contribución específica de cada predictor es el análisis de la comunalidad (Cooley y Lohnes, 1976).

Este método permite determinar la proporción de varianza explicada de la variable dependiente asociada únicamente con cada variable independiente (Rowell, 1996). Por ejemplo, en el caso de ingresar dos predictores, el análisis de la comunalidad divide en tres el porcentaje de varianza explicada: la *varianza específica* de cada predictor y la *varianza común* entre los dos.

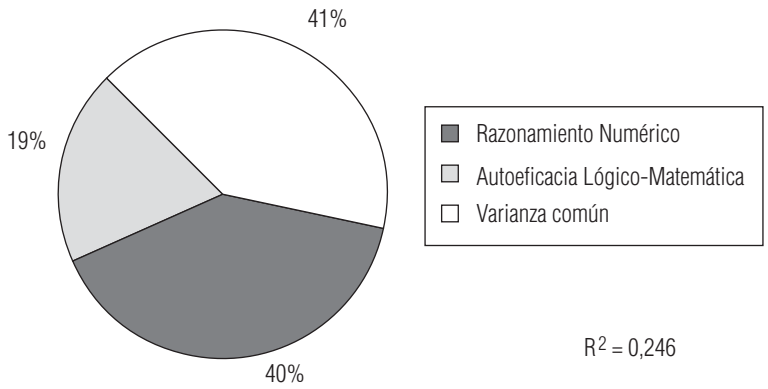
Podemos determinar la varianza única y compartida para las variables independientes a través de cálculos simples: elevando al cuadrado la correlación semiparcial entre el predictor y el criterio o dividiendo la correlación parcial sobre el R^2 (Cooley y Lohnes, 1976). Para obtener la varianza común se resta a la varianza total explicada (R^2) la varianza específica de cada uno de los predictores.

En el ejemplo, Razonamiento Numérico y Autoeficacia Lógico-Matemática explican en conjunto un 25% (R^2 = 0,246) de la variabilidad del promedio en Matemática (véase tabla 9.8). Elevando al cuadrado la correlación semiparcial (313^2) de Razonamiento Numérico y el criterio obtenemos la varianza específica explicada por aquel predictor (0,098). Del mismo modo podemos obtener la varianza específica explicada por Autoeficacia Lógico-Matemática ($0,217^2$ = 0,047). La varianza común es obtenida restando a la varianza total explicada (R^2) la varianza específica de cada uno de los predictores (0,246 – 0,145 = 0,101). Recapitulando, Razonamiento Numérico explica un 10%, Autoeficacia Lógico-Matemática un 5%, y la varianza compartida entre las dos variables independientes explica un 10% de la variabilidad del rendimiento académico en la asignatura Matemática.

Para poder interpretar más acabadamente el análisis de la comunalidad, resulta útil visualizarlo en un gráfico. El total de la varianza explicada por los dos predictores del ejemplo, el R^2, puede representarse como un 100%, y de esta forma dividirse en tres porcentajes, el 40% explicado por la variable Razonamiento Numérico (varianza específica del primer predictor), el 19% explicado por Autoeficacia Lógico-Matemática (varianza específica del segundo predictor) y, por último, un 41% explicado por ambos predictores en común (varianza común).

Este útil recurso gráfico no debería confundir la interpretación estricta de los resultados, es decir, las dos variables independientes de nuestro modelo teórico explican una cuarta parte de la varianza del criterio (rendimiento académico en matemática) y el 75% restante debería atribuirse a la contribución de otras variables no contempladas en el modelo (metas de rendimiento, expectativas de resultados en matemática, responsabilidad en las tareas escolares, aspiraciones educacionales de la familia, autoeficacia para el aprendizaje, entre otras). Cabe aclarar que el análisis de la comunalidad también puede ser utilizado con más de dos predictores.

Figura 4. Porcentaje de varianza específica y común explicada por los predictores del modelo



REFERENCIAS BIBLIOGRÁFICAS

Abad, F.; Garrido, J.; Olea, J. y Ponsoda, V. (2006). Introducción a la psicometría. Teoría Clásica de los Tests y Teoría de Respuesta al Ítem. Madrid: Universidad Autónoma de Madrid. Inédito.

Abelson, R. (1998). *La estadística razonada*. Buenos Aires: Paidós.

Adorno, T.; Frenkel-Brunswik, E.; Levinson, D. y Sanford, R. (1950). *The authoritarian personality*. Nueva York: Harper & Row.

Aftanas, M. S. (1988). Theories, models, and standard systems of measurement. *Applied Psychological Measurement,* 12, 325-338.

Aiken, A. (2003). *Tests psicológicos y evaluación*. México: Prentice Hall.

Allegri, R. F; Mangone, C. A.; Fernández, A.; Rymberg, S.; Taragano, F. E. y Baumann, D. (1997). Spanish Boston Naming Test norms. *The Clinical Neuropsychologist*, 11, 416-420.

American Psychological Association (APA) (1999). *Standards for psychological and educational tests*. Author: Washington, D.C.

Anastasi, A. y Urbina, S. (1998). *Tests psicológicos*. México: Prentice Hall Latinoamericana.

Angoff, W. (1988). Validity: An envolving concept. *Applied Measurement in Education,* 1, 215-222.

Aron, A. y Aron, E. (2001). *Estadística para psicología*. Buenos Aires: Prentice Hall.

Axelrod, B.N.; Aharon-Peretz, J.; Tomer, R. y Fisher, T. (2000). Creating interpretation guidelines for the Hebrew Trail Making Test. *Applied Neuropsychology*, 7, 186-188.

Baker, F. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Education: College Park, MD.

Baldo, M. (2001). Normas regionales y análisis de ítems del test de inteligencia para niños de Wechsler (WISC-III). *Evaluar*, 1, 29-52.

Bandura, A. (1987). *Pensamiento y acción*. Madrid: Martínez Roca.
Bandura, A. (1997). *Self-efficacy: The exercise of control*. Nueva York: Freeman.
Bandura, A. (2001). Guía para la construcción de escalas de autoeficacia. *Evaluar,* 2, 7-38.
Bayley, N. (1993). *Manual of the Bayley Scales of Infant Development* (2nd edition). San Antonio, TX: The Psychological Corporation.
Beck, S.; Steer, R. y Brown, G. (1996). *Beck Depression Inventory-II. Manual.* San Antonio, TX: The Psychological Corporation.
Bem, S. (1974). The measurement of psychological androgyny. *Journal of Consulting & Clinical Psychology*, 42, 165-172.
Bennet, G.; Seashore, H. y Wesman, A. (2000). *Tests de Aptitudes Diferenciales (DAT-5). Manual.* Madrid: TEA Ediciones.
Binet, A. y Simon, T. (1916). *The development of intelligence in children.* Baltimore: Williams & Wilkins.
Bloom, B. (1966). *Taxonomy of educational objectives. The classification of educational goals*. Nueva York: McKay.
Bogardus, E. (1925). Measuring social distances. *Journal of Applied Sociology*, 9, 299-308.
Bond, L. (1996). Norm and criterion-referenced testing. *Practical Assessment, Research & Evaluation, 2.* <www.ericae.net/pare/getvn.asp>. Documento recuperado el 13/11/99.
Briggs, S. R. y Cheek, J. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*, 54, 106-148.
Bunge, M. (1983*). La investigación científica*. Barcelona: Ariel.
Bunge, M. y Ardila, R. (2002). *Filosofía de la psicología*. Barcelona: Ariel.
Butcher, J.; Dahlstrom, W.; Graham, J.; Telegen, A. y Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2 (MMPI-2). Manual for administration and scoring.* Minneapolis: University of Minnesota Press.
Caballo, V. (1987). Evaluación y entrenamiento de las habilidades sociales: una estrategia multimodal. Tesis doctoral (inédita). Universidad Autónoma de Madrid.
Campbell, N. R. (1938). Measurement and its importance for philosophy. *Symposium of the Aristotelian Society*. Londres: Harrison.
Campbell, D. T. y Fiske, A. W. (1954). Convergent and discriminant validation by the multirait-multimethod matrix. *Psychological Bulletin,* 56, 81-105.

Campbell, D. P. y Hansen, J. C. (1981). *Manual for the Strong-Campbell Interest Inventory*. Stanford, CA: Stanford University Press.
Carpenter, P.; Just, M. y Shell, P. (1990). What one intelligence test measures: a theoretical account of the proccesing in the Raven Progressive Matrices Test. *Annual Psychological Review,* 97, 404-431.
Carretero-Dios, H. y Pérez, C. (2005). Normas para el desarrollo y revisión de estudios instrumentales. *International Journal of Clinical and Health Psychology*, 5, 521-551.
Carroll, J. (1993). *Human cognitive abilities*. Londres: Cambridge University Press.
Carrow-Woolfok, E. (1985). *Test for Auditory Comprehension of Language-Revised.* Allen (TX)*:* DLM Teaching Resources.
Carver, C. y Scheier, M. (1996). *Perspectives on personality*. Needham Heights, MA: Allyn & Bacon.
Casullo, M.; Cayssials, A.; Liporace, M.; De Diuk, L.; Arce Michel, J. y Álvarez, L. (1994). *Proyecto de vida y decisión vocacional*. Buenos Aires: Paidós.
Cattell, R. (1966). The Scree Test for the number of factors. *Multivariate Behavioral Research*, 1, 141-161.
Cattell, R. (1967). The theory of fluid and crystallized intelligence. *British Journal of Educational Psychology*, 37, 209-224.
Charter, R. (1999). Sample size requeriments for precise estimates of reliability, generalizability and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, 21, 559-566.
Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20, 37-46.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, Nueva York: Lawrence Earlbaum Associates.
Cohen, R. y Swerdlik, M. (2000). *Pruebas y evaluación psicológicas. Introducción a las pruebas y a la medición*. México: McGraw Hill.
Cooley, W. y Lohnes, P. R. (1976). *Evaluation research in education*. Nueva York: Willey.
Cortada de Kohan, N. (1998). La teoría de la respuesta al ítem y su aplicación al test verbal Buenos Aires. *Interdisciplinaria*, 15, 101-129.
Cortada de Kohan, N. (1999).*Teorías psicométricas y construcción de tests*. Buenos Aires: Lugar Editorial.
Cortina, J. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 96-104.

Costa, P. y Mc Crae, R. (1999). *NEO-PI-R. Manual*. Madrid: TEA ediciones.

Costello, A. y Osborne, J. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation,* 7, 1-9.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika,* 16, 297-334.

Cronbach, L. (1998). *Fundamentos de la evaluación psicológica*. Madrid: Biblioteca Nueva.

Cronbach, L. y Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Cronbach, L. y Glaser, G. (1972). *The dependability of behavioral measurements. Theory of generalizability for scores and profiles*. Nueva York: Wiley.

El Hassan y Jammal (2005). Validation for the test for auditory comprehension of language-revised (TACL-R) in Lebannon. *Assessment in Education*, 2, 183-202.

Embretson, S. y Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Eysenck, H. J. (ed.). (1981). *A model for personality*. Nueva York: Springer.

Eysenck, H. J. y Eysenck, S. B. (1997). *Cuestionario revisado de Personalidad de Eysenck. EPQ-R. Manual*. Madrid: TEA.

Exner, J. (1993). *The Rorschach: A comprehensive system. Volume 1: Basic foundations.* Nueva York: Wiley.

Fernández, A. L; Marino, J. C.; Villacorta, L. y Pérez, E. (2000). Uso y desarrollo de tests en Argentina. Facultad de Psicología. Universidad Nacional de Córdoba, Argentina. Inédito.

Fernández, A. L. y Scheffel, D. L. (2003). A study on the criterion validity of the Mattis Dementia Rating Scale. *International Journal of Testing*, 3, 49-58.

Fernández, A. L. y Marcopulos, B. (2004). About the influence of culture in neuropsychological testing: the case of the Trail Making Test. En revisión.

Ferreyra, R. (1982). Nuevas modalidades para la evaluación educativa. Pruebas referidas a criterio. Facultad de Filosofía y Humanidades, Universidad Nacional de Córdoba. Inédito.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 30, 469-479.

Fogliatto, H. (1991). *Cuestionario de Intereses Profesionales (CIP). Manual*. Buenos Aires: Guadalupe.

Fogliatto, H. y Pérez, E. (2003). Sistema de Orientación Vocacional Informatizado. *SOVI 3. Manual*. Buenos Aires: Paidós.

Folstein, M.; Folstein, S. y McHugh, P. (1975). "Mini-Mental State": a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatry Research*, 19, 189-198.

Galibert, M. S.; Aguerri, M. E.; Lozzia, G. y Abal, F. J. (2006). Análisis del DIF en una escala de voluntad de trabajo mediante los procedimientos de Mantel-Haenszel y Breslow-Day, regresión logística y el criterio ETS. *Actas de las Jornadas de Investigación de la Facultad de Psicología*. Buenos Aires: Universidad Nacional de Buenos Aires.

Gardner, H. (1994). *Estructuras de la mente. La Teoría de las Inteligencias Múltiples*. México: Fondo de Cultura Económica.

Gardner, H. (1999). *Intelligence Reframed*. Nueva York: Basic Books.

Gardner, R. C. (2003). *Estadística para psicólogos usando SPSS para Windows*. México: Prentice Hall.

Gesell, A. y Amatruda, C. (1971). *Diagnóstico del desarrollo normal y anormal del niño*. Buenos Aires: Paidós.

Glaser, R. (1963) Instructional technology and the measurement of learning outcomes: some questions. *American Psychologist*. 18, 519-521.

Glass, G.; McGaw, B. y Smith, M. (1981). *Meta-analysis in Social Research*. California: Sage Publications.

Glutting, J. (2002). Some psychometric properties of a system to measure ADHD among college students. *Measurement and Evaluation in Counseling and Development*, 34, 194-209.

Goldberg, L. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. En M. Mervielde, I.; Deary, F.; De Fruyt, E. y Ostendorf, F. (eds.), *Personality Psychology in Europe, vol.* 7 (pp. 7-28). Tilburg, Holanda: Tilburg University Press.

Gómez Benito, J. (1987). *Meta-análisis*. Barcelona: P.P.V.

Grasso, L. (1969). Posibilidad de prever el rendimiento académico a través de una evaluación de conocimientos previos en la Universidad Nacional de Córdoba. Tesis de Licenciatura en Psicología. Facultad de Filosofía y Humanidades. Universidad Nacional de Córdoba. Inédita.

Grasso, L. (1984). Una escala para la medición de la autoestima en ancianos. *Nueva Revista de Psicología*, 1 (1), 49-62.

Grasso, L. (1999). Introducción a la estadística en Ciencias Sociales y del Comportamiento. Facultad de Filosofía y Humanidades, Universidad Nacional de Córdoba. Inédito.

Guilford, J. (1967). *The nature of human intelligence*. Nueva York: McGraw-Hill.

Gulliksen, H. (1950). *Theories of mental tests*. Nueva York: Wiley.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229-244.

Hambleton, R. K. y Rogers, H. J. (1991). *Fundamentals of item response theory*. Beverly Hills, CA: Sage.

Heaton, R. K.; Chelune, G. J.; Talley, J. L.; Kay, G. G. y Curtiss, G. (1991). *Wisconsin Card Sorting Test Manual*. Odessa, Fl: Psychological Assessment Resources.

Heaton, R.; Grant, I. y Matthews, C. (1991). *Comprehensive norms for an expanded Halstead-Reitan battery: Demographic corrections, research findings, and clinical applications*. Odessa, Fl: Psychological Assessment Resources.

Herrera Rojas, A. (1998). Notas sobre Psicometría. Bogotá: Universidad de Colombia. Inédito.

Himmel, E. (1979). Tendencias actuales en la evaluación del rendimiento escolar. *Revista de Tecnología Educativa,* 2, 3-7.

Ho, D. (1998). Indegenous psychologies. Asian perspectives. *Journal of Cross Cultural Psychology*, 29, 88-103.

Hogan, T. (2004). *Pruebas psicológicas: una introducción práctica*. México: El Manual Moderno.

Holland, J. (1994). *Self-Directed Search. Manual*. Odessa, FL: Psychological Assessment Resources.

Holland, J. (1997). *Making vocational choices*. Englewood Cliffs, NJ: Prentice-Hall.

Holland, P. W. y Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. En H. Wainer y H. I. Braun (eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hood, A. y Johnson, R. (2002). *Assessment in Counseling*. Alexandria, VA: American Counseling Association.

Horn, J. (1965). A rationale and test for the number the factors in factor analysis. *Psychometrika,* 30, 179-185.

Jensen, A. (1998). *The g factor. The science of mental ability*. Westport: Praeger.

Johnson, W. y Bouchard, T. (en prensa). The structure of human intelligence: It is verbal, perceptual and image rotation (VPR), not fluid and crystallized. *Intelligence.*

Johnson, S. (1994). Scholastic Assessment Tests (SAT). En R. Sternberg (ed.), *Encyclopedia of human intelligence* (pp. 956-960). Nueva York: Macmillan.

Juan-Espinosa, M. (1997). *Geografía de la inteligencia humana*. Madrid: Pirámide.

Kahn, J. H. (2006). Factor analysis in Counseling Psychology research, training and practice: Principles, advances and applications. *The Counseling Psychologist*, 34, 1-36.

Kaplan, R. y Saccuzzo, D. (2006). *Pruebas psicológicas. Principios, aplicaciones y temas*. México: Thompson Learning.

Kehoe, (1995). Writing multiple-choice test items. *Practical Assessment, Research & Evaluation,* 7. <www.ericae.net/pare> Documento recuperado el 11/13/99.

Kelly, J. (1987). *Entrenamiento de las habilidades sociales*. Madrid: Desclee de Brower.

Kerlinger, F. y Lee, H. (2002). *Investigación del comportamiento. Métodos de investigación en las ciencias sociales*. México: McGraw-Hill.

Kline, P. (2000). *Handbook of Psychological Testing*. Londres: Routledge.

Kolb, B. y Wishaw, I. (1986). *Fundamentos de neuropsicología humana*. Barcelona: Labor.

Kuder, E. y Zitowski, G. (1991). *Kuder Occupational Interest Survey. General Manual*. Adel, IA: National Career Assessment Services.

Lau, C.W. y Hoosain, R. (1999). Working memory and language difference in sound duration: a comparison of mental arithmetic in Chinese, Japanese, and English. *Psychologia. An International Journal of Psychology in the Orient*, 42, 139-144.

Ledesma, R. y Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: an easy-to-use computer program for carrying out Parallel Analysis. *Practical Assessment, Research & Evaluation*, 12 (2), 1-11.

Lent, R. (2004). Toward a unifying theoretical and practical perspective of well-being and psychosocial adjustement. *Journal of Counseling Psychology*, 51, 482-509.

Lent, R.; Brown, S. y Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest: Choice and performance. *Journal of Vocational Behavior*, 45, 79-122.

Lezak, M. D. (1995). *Neuropsychological assessment*. Nueva York: Oxford University Press.

Likert, R. A. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140.

Loelhin, J. C. (1992). *Genes and environment in personality development.* Nuevabury Park, CA: Sage.
Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Lawrence Erlbaum Associates.
Lord, F. y Novick, M. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.
Mac Combs, B. L. y Branan, L. (1990). *Social skills for job success: Teacher's guide*. Baltimore, MD: Educational Press.
Mantel, N. y Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of The National Cancer Institute*, 22, 719-748.
Martínez Arias, R. (1995). *Psicometría*. Madrid: Síntesis Psicológica.
Mc Crae, R. R; Yik, M. S. M.; Trapnell, P. D.; Bond, M. H. y Paulhus, D. L. (1998). Interpreting personality profiles across cultures: bilingual, acculturation, and peer rating studies of chinese students. *Journal of Personality and Social Psychology*, 74, 1041-1055.
McGrew, K.; Flanagan, D.; Keith, T. y Vanderwood, M. (1997). Beyond g: the impact of Gf-Gc especific cognitive abilities research on the future use and intepretation of intelligence tests in the school. *School Psychology Review*, 26, 189-210.
Merenda, P. (1997). A guide to the proper use of Factor Analysis in the conduct and reporting of research: pitfalls to avoid. *Measurement and Evaluation in Counseling and Evaluation,* 30, 156-163.
Mitchell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale: Erlbaum.
Moreno, R.; Martínez, R. y Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema,* 6, 3, 490-497.
Multon, R.; Brown, S. y Lent, R. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology,* 38, 30-38.
Muñiz, J. (1998). La medición de lo psicológico. *Psicothema,*10, 1, 1-21.
Muñiz, J. (2001). *Teoría Clásica de los Tests*. Madrid: Pirámide.
Muraki, E. (1993). *POLYFACT* (computer program). Princeton, NY: Educational Testing Service.
Murat, F. (1984). Actitud hacia la matemática. *Nueva Revista de Psicología*, 1, 13-25.
Murat, F. (1985). *Evaluación del comportamiento humano.* Universidad Nacional de Córdoba.

Norman, W. (1963). Toward an adequate taxonomy of personality attributes. *Journal of Abnormal and Social Psychology,* 65, 574-583.
Nunnally, J. (1991). *Teoría psicométrica* (2ª ed.). Buenos Aires: Paidós.
Nunnally, J. y Bernstein, I. (1995). *Teoría psicométrica*. México: Mc Graw Hill.
Oesterlind, S. (1990). Establishing criteria for meritorius test items. *Educational Research Quality*, 3, 26-30.
Olea, J.; Ponsoda, V. y Prieto, G. (eds.) (1999). *Tests adaptativos informatizados.* Madrid: Pirámide
Oliden, P. E. (2003). Sobre la validez de los tests. *Psicothema*, 15, 315-321.
Orwin, R. G. (1983). A fail safe N for effect size. *Journal of Educational Statistics*, 8, 157-159.
Padua, J. (1979). *Técnicas de investigación aplicadas a las Ciencias Sociales.* México: Fondo de Cultura Económica.
Pajares, F.; Hartley, J. y Valiante, G. (2001). Response format in Writing Self-Efficacy Scales. Greater discrimination increases prediction. *Measurement and Evaluation in Counseling and Development,* 33, 214-221.
Pérez, E. (2001). Construcción de un Inventario de Autoeficacia para Inteligencias Múltiples. Tesis Doctoral. Córdoba, Argentina: Universidad Nacional de Córdoba. Inédita.
Pérez, E. y Gay, D. (1991). La utilización de pruebas psicológicas en la ciudad de Córdoba. Facultad de Psicología. Universidad Nacional de Córdoba, Argentina. Inédito.
Pérez, E.; Cupani, M. y Ayllón, S. (2005). Predictores de rendimiento académico en la escuela media: habilidades, autoeficacia y rasgos de personalidad. *Avaliação Psicológica*, 4, 1, 1-12.
Plomin, R.; DeFries, J.; McClearn, G. y McGuffin, P. (2002). *Genética de la conducta*. Barcelona: Ariel Ciencia.
Popham, W. J. (1975): *Educational evaluation*. Englewood Cliffs, NJ: Prentice Hall.
Prieto, G. y Delgado, A. (1999). Medición cognitiva de las aptitudes. En Olea, V.; Ponsoda, J. y Prieto, G. (eds.). *Tests informatizados. Fundamentos y aplicaciones* (pp. 28-57). Madrid: Pirámide.
RASCAL (1989). *Rasch Item Calibration Program*. St. Paul: Assessment System Corporation.
Rasch, G. (1963). *Probabilistic models for some intelligence and attainment tests*. Copenhague: Denmarks Paedagogiske Institut.
Raven, J. (1993). *Test de Matrices Progresivas*. *Manua*l. Buenos Aires: Paidós.

Reise, S.; Waller, N. y Comrey, A. (2000). Factor analysis and scale revision. *Practical Assessment*, 12, 287-297.

Rey, A. (1941). L'examen psychologique dans les cas d'encèphalopathie traumatique. *Archives de Psychologie*, 28, 286-340.

Richaud de Minzi, C. (2005). Estilos parentales y afrontamiento en niños. *Revista Latinoamericana de Psicología*, 37, 1, 47-58.

Rorschach, H. (1921). *Psychodiagnostik*. Berne, Switzerland: Bircher.

Rosenthal, R. y Di Matteo, M. R. (2001). Meta-Analysis: Recent Developments in Quantitative Methods for Literature Reviews. *Annual Reviews Psychology*, 52, 59-82 .

Rosenthal, R. y Di Matteo, M. R. (2002). Meta-Analysis. En H. Pashler *et al.* (eds.). *Handbook of Experimental Psychology* (3ª ed.) (pp.391- 428). Nueva York: John Wiley & Sons.

Roth, G. y Dicke, U. (2005). Evolution of the brain and intelligence. *Trends in Cognitive Sciences*, 9, 250-257.

Rowell, R. K. (1996). Partitioning predicted variance into constituent pars: How to conduct regression commonality analysis. En Thompson, B. (ed.). *Advances in social science methodology*, vol. 4 (pp. 33-43). Greenwich, CT: JAI Press.

Russell, M. T. y Karol, D. L. (2000). *16PF-5. Manual.* (5ª ed.) Madrid: TEA.

Samejina, F. (1973). Homogeneous case of the continous response model. *Psychometrika*, 38, 203-219.

Santiesteban, C. (1990). *Psicometría*. Madrid: Norma.

Shearer, B. (1999). *Multiple Intelligence Developmental Assessment Scale (MIDAS). Professional Manual*. Kent: University of Kent Press.

Spearman, C. (1927). *The nature of "intelligence" and the principles of cognition*. Londres: MacMillan.

Spielberger, C. (1983). *State-Trait Anxiety Inventory for Adults. Manual.* Redwood City, CA: Mind Garden.

Statistical Package for the Social Sciences (SPSS) (1995). *Statistical Package for the Social Sciences Reference Guide*. Chicago: Autor.

Sternberg, R. (1987). *Inteligencia humana*. Vol. I. Barcelona: Paidós.

Stevens, S. (1949). On the theory of scales of mesurement. *Science,* 103, 677-680.

Stroop, J. R. (1935). Studies of interference in serial verbal reaction. *Journal of Experimental Psychology,* 18, 643-662.

Tabachnick, B. y Fidell, L. (2001). *Using multivariate statistics*. Nueva York: Harper & Row.

Tanzer, N. K. (1995). Cross-cultural bias in likert-type inventories: perfect matching structures and still biased? *European Journal of Psychological Assessment*, 11, 194-201.

Taub, G. (2001). A confirmatory analysis of the Wechsler adult intelligence scale-third edition: is the verbal-performance discrepancy justified? *Practical Assessment, Research and Evaluation*, 6, 1-11.

Tellegen, A. (1988). Personality similarity in twins reared apart and together. *Journal of Personality and Social Psychology*, 54, 1031-1039.

Tetreau, B. y Trahan, M. (1986). *Test Visuel d´Interets Tetreau-Trahan. Manuel d´usage.* Montreal: SECOROP.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis*. Washington, DC: American Psychological Association.

Thompson, B. y Borrello, G. (1985). The importance of structure coefficients in regression research. *Educational and Psychological Measurement*, 28, 203-209.

Thompson, B. y Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: An historical overview and some guidelines. *Educational and Psychological Measurement,* 56, 197-208.

Thorndike, R. L. (1989). *Psicometría aplicada*. México: Limusa.

Thurstone, L. L. (1935). *The vectors of the mind*. Chicago: University of Chicago Press.

Tokar, D. M.; Fischer, A. R. y Subich, L. M. (1998). Personality and vocational behavior: A selective review of the literature, 1993-1997. *Journal of Vocational Behavior*, 53, 115-153.

Tornimbeni, S. y González, C. (1997). *Construcción de una escala de actitudes ante la investigación en Psicología*. Facultad de Psicología. Universidad Nacional de Córdoba. Inédito.

Tracey, T. (2002). Personal Globe Inventory: measurement of the spherical model of interests and competence beliefs. *Journal of Vocational Behavior*, 60, 113-172.

Tyler, R. (1978). *The Florida Accountability Program*. Washington DC: National Education Association.

Van de Vijver, F. J. F. y Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Nuevabury Park, CA: Sage.

Van de Vijver, F. J. R. y Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: an overview. *European Review of Applied Psychology*, 47, 263-279.

Van der Lienden, W. y Hambleton, R. (1997). *Handbook of modern Item Response Theory*. Nueva York: Springer.

Velandrino, A. (1998). *Análisis de datos en ciencias sociales*. Murcia: DM Editora.
Vernon, P. (1964). *The structure of human abilities*. Londres: Metluen.
Walsh, W. y Betz, N. (1990). *Tests and assessment*. Englewood Cliffs, NJ: Prentice Hall.
Wechsler, D. (1994). *Test de Inteligencia para niños (WISC-III). Manual Técnico.* Buenos Aires: Paidós.
Wechsler, D. (1999). *WAIS-III. Test de Inteligencia para adultos. Manual Técnico.* Buenos Aires: Paidós.
Wechsler, D. (2005). WISC-IV. *Escala de Inteligencia de Wechsler para niños. Manual.* Madrid: TEA.
Wissler, C. (1901). *The correlation of mental and physical tests.* Nueva York: Columbia University.
Wolf, F. (1986). *Meta-analysis: Quantitative methods of research synthesis*. Beverly-Hills, CA: Sage.
Woodcock, R.; McGrew, K. y Mather, N. (2001). *Woodcock-Johnson III. Tests of Cognitive Abilities.* Itasca, IL: Riverside Publishing.
Woolfolk, A. (2006). *Psicología educativa*. México: Pearson Educación.
Yerkes, R. (1921). Psychological examining in the United States Army. *Memoirs of the National Academy of Scienc*e, vol. 15.