5

Concepto y evidencias de validez

El concepto de validez

Como ya hemos visto en los capítulos precedentes, los estudios de fiabilidad nos informan de si estamos midiendo con mucho o con poco error de medida, pero no informan de qué atributo estamos midiendo. Los estudios de validez van a aportar significado a las puntuaciones que estamos obteniendo, permitiéndonos conocer si el uso que pretendemos hacer de ellas es o no adecuado.

Los indicadores de fiabilidad son diferentes en las distintas teorías y el concepto de fiabilidad sólo ha ido matizándose a lo largo de los años, ligado al desarrollo de las distintas teorías de los tests. El concepto de validez, sin embargo, ha cambiado enormemente, tal como veremos en el último apartado del presente capítulo. La validez ha sido definida de muchas maneras a lo largo de la historia de la Psicometría y su definición sigue cambiando. Lo único que se ha mantenido a través del tiempo es su consideración como la propiedad más importante y fundamental al desarrollar y evaluar un test (p.ej.: Anastasi y Urbina, 1997; Cronbach, 1988).

La conceptualización actualmente dominante sobre la validez es la que recogen los *Standards for Educational and Psychological Testing* en su edición de 1999. En este documento, elaborado por tres importantes asociaciones profesionales americanas (AERA, APA y NCME), se define la validez como el grado en que la teoría y los datos disponibles apoyan la interpretación de las puntuaciones de un test para un uso concreto. Hay varios aspectos a destacar de esta definición:

1. Al igual que ocurre en el caso de la fiabilidad, ya no hablamos de validez de un test sino de validez de las puntuaciones de un test. No tiene sentido hablar de "propiedades del test", ya que éstas dependen del contexto de evaluación y de la muestra.

2. El profesional responsable de la aplicación de un test debe consultar el manual del mismo para averiguar si la utilización e interpretación pretendida por él coincide con la proporcionada por la documentación de la prueba. En caso negativo, para poder realizar la interpretación pretendida deberá desarrollar una investigación, un estudio de *validación*, que le permita recoger información que apoye (o no) su utilización. Por lo tanto, la validación no solo incumbe a los tests de nueva creación, sino que representa un proceso de acumulación de evidencias que apoyan las interpretaciones propuestas para las puntuaciones de un test, para así lograr una mejor comprensión de su significado.

Siguiendo la propuesta de Kane (2006a), que es consistente con la visión de los Standards, el proceso de validación implicaría el uso de dos argumentos: el argumento interpretativo y el argumento de validez. El proceso de validación debe comenzar con el desarrollo del *argumento interpretativo*, que supone proponer con detalle interpretaciones y usos de las puntuaciones. Por ejemplo, debemos especificar todas las asunciones en las que se basa el test, los componentes del constructo, las diferencias con otros constructos y sus relaciones con otras variables. Si las interpretaciones y usos no están claramente identificados entonces no pueden ser evaluados.

El segundo, el *argumento de validez*, consiste en evaluar el argumento interpretativo. La interpretación propuesta para las puntuaciones determina las clases de evidencia necesarias para la validación. Es posible que una o varias de las interpretaciones sean válidas mientras que otras se consideren inválidas. Por ejemplo, es posible que un test de personalidad sea adecuado para un proceso de selección de personal, pero no lo sea para un proceso de diagnóstico de patologías. El argumento de validez implica la evaluación de las interpretaciones propuestas a través de una serie de análisis lógicos y estudios empíricos, siendo siempre necesaria la integración de las diferentes clases de evidencia. Las *evidencias de validez* son, por lo tanto, las pruebas recogidas para apoyar la interpretación propuesta. La principal ventaja de esta aproximación sería que intenta proporcionar una guía para dirigir los esfuerzos de investigación. Las clases de evidencia que serían más relevantes son aquellas que apoyan inferencias y asunciones del argumento interpretativo que son más problemáticas. Además, la etapa de evaluación también implica una búsqueda de asunciones ocultas y de posibles interpretaciones alternativas de las puntuaciones.

Si concebimos la validación como el proceso investigador en el que se van acumulando evidencias sobre la interpretación de las puntuaciones de un test, resulta patente que para obtener estas evidencias podremos usar una enorme variedad de métodos o estrategias. De ahí que ya no usemos el término "tipos de validez" sino el de "tipos de evidencia", intentando resaltar el carácter unitario del concepto de validez. Estas diferentes fuentes de evidencia no representan distintos tipos de validez. Ahora se plantea el estudio de evidencias basadas en: el contenido, la estructura interna, la relación con otras variables, el proceso de respuesta, y las consecuencias de la aplicación del test.

Para analizar los datos de las distintas fuentes de evidencia se usan una amplia variedad de técnicas, que por su importancia y especificidad se tratarán en distintos capítulos específicos. Concretamente, para obtener evidencias relativas a la estructura interna de las puntuaciones es preponderante el uso del Análisis Factorial Exploratorio (AFE) y del Análisis Factorial Confirmatorio (AFC). Estas dos técnicas se exponen, respectivamente, en los capítulos 6 y 10. Dentro de las evidencias relativas a la estructura interna también pueden ubicarse los trabajos encaminados a evaluar el funcionamiento diferencial de los

ítems (FDI); la definición y la tecnología para la detección del FDI se proporciona en el capítulo 13. Por otra parte, en el capítulo 14, se incluyen otros procedimientos que se aplican para obtener información sobre la relación del test con otras variables (p. ej.: la regresión lineal múltiple) y sobre la generalización de la validez.

Hay numerosos manuales en los que se trata el concepto y las evidencias de validez, tanto en castellano (p.ej.: Martinez Arias, Hernandez-Lloreda y Hernandez-Lloreda, 2006; Muñiz, 2002; Navas, 2001), como en inglés (p.ej.: Carmines y Zeller, 1979; Crocker y Algina, 1986; Wainer y Braun, 1988).

Evidencias basadas en el contenido del test

Definición

Es fácil comprender la necesidad de examinar el contenido de un test como un primer paso para juzgar si un instrumento puede usarse para un propósito en particular. Por ejemplo, estudiantes, padres y profesores esperan que las preguntas de un examen de Lengua de Educación Primaria sean consistentes con los objetivos curriculares para esa asignatura y nivel. Esto es especialmente evidente en los tests educativos. No sorprende, por tanto, que la necesidad de examinar el contenido de los tests apareciese ya en 1954 en un documento de la APA sobre recomendaciones técnicas para el diseño y uso de los tests.

Es necesario aclarar que por contenido del test no nos referimos únicamente a los ítems que lo componen. Actualmente se incluyen, además, las instrucciones para su administración y las rúbricas o criterios para su corrección y puntuación.

Sireci (2003) indica que hay al menos dos aspectos esenciales a tener en cuenta para realizar la validación de contenido: la definición del dominio, y la representación del dominio. La *definición del dominio* se refiere a la definición operativa del contenido. En la mayoría de los tests educativos esta definición tiene la forma de una tabla de especificaciones de doble entrada, en la que las filas indican las áreas de contenido relevantes para el dominio en cuestión y las columnas indican las operaciones o procesos cognitivos implicados en la resolución de las tareas planteadas. Se especifican además los porcentajes de ítems asignados a cada combinación de área y proceso cognitivo.

Las empresas de tests más importantes de USA. (p.ej.: American College Testing, California Bureau Test, Educational Testing Service…) suelen emplear estas tablas. Así por ejemplo, en la página Web del National Assessment of Educational Program (NAEP)[1] podemos encontrar varios ejemplos. El NAEP es el programa de evaluación del rendimiento académico llevado a cabo por el Departamento de Educación de USA, que permite comparar el rendimiento de los estudiantes en la escuela en varias materias y en todos los estados. Si tomamos, por ejemplo, la tabla de especificaciones para la evaluación del progreso educativo en Geografía vemos que incluye tres áreas de contenidos: espacio y lugar, ambiente y sociedad y conexiones y dinámicas espaciales. Las dimensiones cognitivas evaluadas son conocimiento, comprensión y aplicación. Se muestran, además, los porcentajes de distribución de ítems por áreas de contenido y algunos ejemplos de ítems para cada combinación de área y habilidad cognitiva. Así por ejemplo, el ítem *"¿Qué factores es-*

[1] http://nces.ed.gov/nationsreportcard/geography/elements.asp

timulan las migraciones humanas?" Está diseñado para medir "conocimiento" en el área "conexiones y dinámicas espaciales". Mientras que el ítem *"Explique las razones que los mexicanos y cubanos tienen hoy en día para emigrar a los Estados Unidos"* está diseñado para medir "comprensión" en la misma área de conocimiento.

Para definir el dominio de manera adecuada podemos usar varias fuentes. En los tests educativos es habitual usar los libros de texto y los objetivos curriculares; en el ámbito de selección de personal es frecuente usar los resultados de los análisis de puestos de trabajo; los datos obtenidos en tales análisis se usan para defender la evaluación de áreas específicas y para establecer su importancia en el test (p.ej., la proporción de ítems de cada una). En los tests de aptitudes se utilizan las teorías sobre las habilidades mentales y su funcionamiento.

Hasta ahora nos hemos ocupado de la definición del dominio. El segundo elemento resaltado por Sireci es la *representación del dominio*, que a su vez abarca dos aspectos: la representatividad y la relevancia. La representatividad o cobertura del dominio indica la adecuación con que el contenido del test representa todas las facetas del dominio definido. Hay que examinar si todo el contenido del dominio está siendo medido y si hay facetas concretas que han sido infrarrepresentadas. Por su parte, al estudiar la relevancia examinamos el grado en que cada ítem del test mide el dominio definido, pudiéndose detectar problemas relativos a la presencia de contenidos irrelevantes.

Procedimientos

La mayoría de los estudios de validación de contenido requieren del trabajo de jueces o expertos que evalúan los ítems del test y emiten juicios sobre el grado de emparejamiento entre los ítems y los objetivos definidos en la tabla de especificaciones. Habitualmente se trabaja con un reducido número de jueces que emiten una cantidad importante de evaluaciones. Es crucial realizar una cuidadosa selección de los expertos. En un estudio "tradicional" de validez de contenido, una vez identificado el grupo de expertos en el dominio evaluado, éstos deben informar del grado en que el dominio está bien definido y del grado en que el test lo representa bien. Se pueden utilizar varios procedimientos para que los jueces evalúen el emparejamiento entre los ítems y los objetivos del test.

Rovinelli y Hambleton (1977) propusieron una tarea en la que cada juez juzga si el contenido de cada ítem refleja cada uno de los objetivos especificados. El juez debe asignar "+1" si considera que el ítem mide el objetivo, "−1" si cree que no lo mide y "0" si tiene dudas sobre si lo mide o no. Por ejemplo, en un test de 10 objetivos y 4 ítems por objetivo cada juez debería realizar 400 juicios. El *índice de congruencia ítem-objetivo* se obtiene mediante la expresión:

$$I_{jk} = \frac{N}{2N-2}\left(\overline{X}_{jk} - \overline{X}_{j}\right) \qquad [5.1]$$

Siendo N el número de objetivos, $\overline{X}_{jk}$ la media de los jueces para el ítem j en el objetivo k y $\overline{X}_{j}$ la media de los jueces para el ítem j en todos los objetivos.

Este índice toma valores entre −1 y 1. Nótese que un valor del índice de 1 en un ítem

indicaría que todos los jueces lo han valorado +1 en el objetivo *k* (la media del ítem en el objetivo sería igual a 1) y −1 en todos los demás objetivos. Un valor del índice de −1 indicaría que todos los jueces lo han valorado —1 en el objetivo *k* y +1 en todos los demás objetivos. Podemos fijar un punto de corte para decidir qué ítems presentan valores adecuados y cuáles no. Por ejemplo, si tenemos 20 jueces y 10 áreas de contenido. Podríamos exigir que al menos 15 jueces valorasen el ítem como adecuado para el objetivo propuesto e inadecuado para los otros. En este ejemplo, el índice debería valer 0,75.

Ejemplo 5.1. El índice de congruencia de Rovinelli y Hambleton

En la Tabla 5.1 se muestran las hipotéticas evaluaciones de una muestra de 10 jueces recogidas con este procedimiento para un test de 6 ítems que pretende medir 2 objetivos. En negrita se muestran los ítems que han sido diseñados para medir cada objetivo. Así, los tres primeros ítems fueron diseñados para evaluar el objetivo 1 y los tres últimos para el objetivo 2. Cada juez debe realizar 12 valoraciones (6 ítems x 2 objetivos). Por ejemplo, el juez 1 evalúa con "+1" al ítem 3 en el objetivo 1 (cree que lo mide) y con "0" a ese mismo ítem en el objetivo 2 (tiene dudas sobre si lo mide o no).

Tabla 5.1. Evaluaciones hipotéticas de 10 jueces para un test de 6 ítems que mide 2 objetivos

		Jueces										
Objetivos	*Ítems*	1	2	3	4	5	6	7	8	9	10	$\sum_{i=1}^{i=10} X_i$
1	**1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**0**	**+1**	**+1**	**+1**	9
	2	**+1**	**+1**	**0**	**+1**	**+1**	**+1**	**0**	**+1**	**+1**	**+1**	8
	3	**+1**	**0**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**−1**	7
	4	−1	0	−1	−1	−1	−1	−1	0	0	−1	−7
	5	−1	0	−1	−1	−1	−1	−1	−1	−1	−1	−9
	6	−1	0	−1	−1	−1	−1	−1	0	−1	−1	−8
2	1	−1	0	−1	−1	−1	−1	−1	0	−1	−1	−8
	2	−1	0	−1	−1	−1	−1	−1	−1	0	−1	−8
	3	0	+1	−1	−1	−1	−1	−1	−1	−1	0	−6
	4	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**0**	**+1**	**+1**	**+1**	9
	5	**+1**	**+1**	**0**	**+1**	**0**	**+1**	**0**	**−1**	**+1**	**+1**	6
	6	**+1**	**+1**	**+1**	**+1**	**+1**	**0**	**+1**	**−1**	**+1**	**0**	6

El índice de congruencia del ítem 3 y el objetivo 1 (que es el que pretende medir) es:

$$I_{jk} = \frac{N}{2N-2}\left(\overline{X}_{jk} - \overline{X}_{j}\right) = \frac{2}{4-2}\left(\frac{7}{10} - \frac{7+(-6)}{20}\right) = 0{,}65$$

Supongamos que para considerar un ítem adecuado decidimos que al menos 7 de los 10 jueces valoren el ítem como apropiado para el objetivo propuesto e inapropiado para el otro objetivo. En este caso, el índice debería ser al menos de 0,7. El índice de congruencia del ítem 3 no alcanza este valor, por lo que no se consideraría adecuado para evaluar el objetivo en cuestión.

Un segundo procedimiento, más sencillo que el anterior, implica el uso de una tarea de emparejamiento. Se presentan a los jueces dos listas, una con los ítems y otra con los objetivos. La tarea del juez consiste en indicar qué objetivo piensa que mide cada ítem (si es que mide alguno). Estas clasificaciones serían usadas para obtener "índices de congruencia ítem-objetivo", así como "índices de congruencia globales" para cada área de contenido. Por ejemplo, si en un examen de Geografía un ítem diseñado para medir "conocimiento" en el área "conexiones y dinámicas espaciales" fuese clasificado en esa categoría por un 80% de los jueces, su índice de congruencia sería de 0,8. Se suele considerar que índices de 0,7 o mayores corresponden a ítems congruentes con su objetivo.

Los índices de congruencia son fáciles de comprender y de calcular y proporcionan información sobre la representación del dominio. Además, a partir de los datos anteriores hay que ver el porcentaje de ítems que hay en cada una de las celdas de la tabla de especificación y ver si éste es acorde con lo propuesto en la definición del dominio. Los datos recogidos de esta manera pueden resumirse usando estadísticos descriptivos como: el % de ítems que se emparejan a los objetivos, la correlación entre el peso dado al objetivo y el número de ítems que miden el objetivo o el porcentaje de objetivos no evaluados por ninguno de los ítems. En ocasiones, para evaluar la precisión con que los jueces llevan a cabo su tarea, se incluyen ítems que no miden ninguno de los objetivos (ítems de relleno). Se evalúa la efectividad de los jueces mediante el recuento del número de tales ítems que no han sido identificados por cada juez. Aquellos jueces que no logren un mínimo nivel de ejecución se eliminan del análisis.

Ejemplo 5.2. Tarea de emparejamiento para evaluar la validez de contenido

La Tabla 5.2 muestra un resumen y análisis de las evaluaciones hipotéticas de 5 jueces en un test de 12 ítems que mide 4 objetivos. Los ítems 13, 14 y 15 no medían ninguno de los objetivos. En los ítems que componen el test, una puntuación de 1 significa que el juez asignó el ítem al objetivo para el que había sido desarrollado. Una puntuación de 0 significa que el juez no asignó el ítem al objetivo para el que fue desarrollado. En los "ítems de relleno" una puntuación de 0 significa que el juez no asignó el ítem a ninguno de los objetivos. Una puntuación de 1 significa que el juez asignó el ítem a alguno de los objetivos. Por ejemplo, el juez nº 2 consideró que el ítem 2 no medía el objetivo 1, pero que el ítem 7 sí medía ese objetivo. Si nos fijamos en los datos del ítem 10 veremos que sólo uno de los cinco jueces consideraron que midiese el objetivo (2) para el que había sido desarrollado, de ahí que su índice de congruencia tenga un valor de 0,2.

Respecto a la efectividad con que los jueces realizan su tarea, vemos que el juez me-

nos eficaz ha sido el nº 2, ya que no detecta ninguno de los ítems "de relleno" introducidos. Este juez es también quien ha realizado un menor porcentaje de clasificaciones "congruentes" (58%). Es decir, de los 12 ítems que componen el test solo emparejó 7 con el objetivo para el que habían sido diseñados. Por lo tanto, este sería un juez cuyos datos deberían, probablemente, ser eliminados.

Tabla 5.2. Hipotética tarea de emparejamiento llevada a cabo por cinco jueces

		Jueces					*Índice de*
Objetivos	*Ítem*	1	2	3	4	5	*congruencia*
1	2	1	0	1	1	1	0,8
	7	1	1	0	1	1	0,8
2	1	1	1	1	1	1	1
	3	1	1	1	1	1	1
	8	1	1	1	1	0	0,8
	10	0	0	0	1	0	0,2
	11	1	0	0	0	1	0,4
3	4	1	1	1	0	0	0,6
	6	1	0	1	0	0	0,4
4	5	0	0	1	1	1	0,6
	9	1	1	1	0	1	0,8
	12	1	1	0	1	1	0,8
% de clasificaciones "congruentes"		83	58	67	67	67	
	13	0	1	0	0	1	
Ítems "de relleno"	14	1	1	0	1	0	
	15	1	1	0	0	0	
Nº ítems "de relleno" no identificados		2	3	0	1	1	

El juicio solicitado a los expertos no tiene por qué ser dicotómico (clasificar un ítem, o no, en un objetivo). Hambleton (1980, 1984) propuso el uso de escalas tipo Likert de 5 puntos donde la tarea del juez es evaluar la relevancia de cada ítem para medir el objetivo pensado. No hay un número estándar de puntos a utilizar. Se suele aconsejar no usar menos de 5 puntos ni más de 9. El valor más bajo de la escala indica "nada relevante" y el más alto "completamente relevante". Se obtienen la media y/o mediana de las valoraciones, que se usan como un *índice de la relevancia del ítem*. La media de los índices de relevancia para todos los ítems de un área de contenido se usa como *índice de representación* de esa área. Es el investigador quien debe decidir cuándo se considera que el resultado obtenido para un ítem es lo suficientemente bueno como para considerarlo relevante; por ejemplo, en

una escala de 5 puntos una media superior a 3,5 o una mediana superior a 3. Además, para cada juez se puede calcular la discrepancia entre su valoración y la mediana para cada ítem. Para ello se deben sumar las diferencias, en valor absoluto, entre la valoración dada por el juez y la mediana para cada ítem. Cuando la discrepancia cometida es importante se puede dudar de la competencia del juez, y por tanto eliminarlo de los análisis.

Ejemplo 5.3. Evaluaciones de una muestra de jueces usando una escala tipo Likert

La Tabla 5.3 incluye las evaluaciones de 10 jueces, usando una escala de 7 puntos, en un test de 14 ítems que pretende medir 2 objetivos. Los 7 primeros ítems medían el primer objetivo y los 7 últimos el segundo. Por ejemplo, el juez nº 2 consideró que el tercer ítem, diseñado para medir el primer objetivo, lo hacía en un grado de 4 sobre 7.

Tabla 5.3. Resumen de las evaluaciones de 10 jueces en un test de 14 ítems usando una escala de 7 puntos

		Jueces												
Objetivos	*Ítems*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*media*	*Mdn*	*Rango*
1	1	7	5	4	5	3	4	4	7	7	5	5,1	5	4
	2	7	6	7	7	3	7	6	7	7	7	6,4	7	4
	3	4	4	1	3	3	2	3	2	5	6	3,3	3	5
	4	1	6	7	1	6	7	3	2	4	7	4,4	5	6
	5	6	6	6	5	4	6	6	6	7	7	5,9	6	3
	6	4	6	7	5	4	7	5	6	7	7	5,8	6	3
	7	7	5	5	6	3	6	4	6	5	6	5,3	5,5	4
2	8	3	4	1	1	4	4	3	5	6	6	3,7	4	5
	9	6	6	2	6	1	4	3	6	7	5	4,6	5,5	6
	10	3	6	3	4	1	4	4	5	7	6	4,3	4	6
	11	7	6	4	6	1	2	6	7	6	5	5,0	6	6
	12	7	3	5	7	1	6	4	7	5	6	5,1	5,5	6
	13	7	5	2	6	1	4	2	7	7	7	4,8	5,5	6
	14	7	6	3	7	4	5	4	7	7	6	5,6	6	4
Discrepancia de cada juez respecto a la Mdn		18	9	23	13	37	14	17	14	17	15			

Si observamos las medias y medianas de los ítems podremos concluir que, según esta muestra de jueces, los ítems que mejor reflejan los objetivos planteados son: para el objetivo 1, el ítem 2 (con una media de 6,4 y una mediana de 7), y para el objetivo 2, el ítem 14 (con una media de 5,6 y una mediana de 6). Si observamos el rango de las evaluaciones obtenidas por un ítem, tendremos un indicador del nivel de desacuerdo de los jueces.

Por ejemplo, los ítems 5 y 6 presentan el menor rango (3), lo que nos informa de un mayor nivel de acuerdo entre los jueces al evaluarlos. Respecto al análisis de las discrepancias de los expertos, el juez nº 5 destaca claramente[2]. Sus evaluaciones son las que mayores diferencias obtienen respecto a las medianas. Por lo que, en este caso, sus valoraciones serían candidatas a ser eliminadas del análisis definitivo.

Obviamente, para que los datos recogidos mediante cualquiera de los procedimientos que hemos detallado anteriormente sean informativos hay que garantizar que existe una adecuada fiabilidad interjueces, es decir, que las valoraciones que realizan son consistentes. En el capítulo 9 se muestran algunos de los múltiples índices que se han propuesto para su estudio. Una limitación de los índices de congruencia y de relevancia que acabamos de describir es que, al informar a los jueces de lo que el test se supone que mide, estamos restringiendo sus evaluaciones a las dimensiones propuestas y, por lo tanto, influenciando sus percepciones sobre lo que mide el ítem. El conocimiento por parte de los jueces de los objetivos del test puede sensibilizarles a las expectativas de los constructores del test y crear un sesgo potencial de demandas de la tarea que contamine sus juicios. Esto probablemente sobreestima los índices de relevancia y congruencia obtenidos. Para superar estos problemas se han propuesto métodos que intentan descubrir las percepciones de los jueces sin informarles de las áreas específicas de contenido del test. Concretamente, Sireci y Geisinger (1992, 1995) utilizaron métodos de escalamiento multidimensional y análisis de conglomerados con los juicios sobre la similaridad del contenido medido por pares de ítems. En este caso, la tarea de los jueces es evaluar, usando una escala tipo Likert, la similaridad entre todos los posibles pares de ítems del test con respecto al conocimiento o habilidades cognitivas medidas. El objetivo era determinar si la estructura propuesta en la tabla de especificaciones era congruente con las evaluaciones de similaridad dadas por los expertos. Tal y como el propio Sireci indica, el método basado en las similaridades complementa, pero no excluye, los métodos tradicionales. Los trabajos de validación de contenido deben incluir los procedimientos tradicionales basados en los índices de congruencia y relevancia.

Como ya dijimos al principio de este apartado, la mayoría de los trabajos de validación de contenido están basados en las evaluaciones de jueces, pero también se ha propuesto el examen del contenido de los tests a partir de las respuestas dadas por los sujetos que responden al mismo. Se han aplicado técnicas de escalamiento multidimensional, y análisis cluster (p. ej.: Deville, 1996; Napior, 1972; Oltman, Stricker y Barrows, 1990), análisis factorial (p. ej.: Dorans y Lauwrence, 1987), la Teoría de la Generalizabilidad (p. ej.: Green 1983) y modelos de ecuaciones estructurales (p. ej.: Ding y Hersberger, 2002). Sireci (1998) es crítico con algunas de estas aplicaciones porque considera que confunden las propiedades de los datos con las interpretaciones del contenido. Sin embargo, considera prometedores los estudios basados en la Teoría de la Generalizabilidad.

---

[2] $37 = |3-5| + |3-7| + ... + |4-6|$

Evidencias basadas en la estructura interna del test

¿Mide nuestro test un constructo coherente o se trata simplemente de un conjunto de ítems no relacionados? Las evidencias sobre la estructura interna nos permitirán responder a esta pregunta. Para analizar la estructura interna del test se realizan estudios sobre la dimensionalidad y sobre el funcionamiento diferencial de los ítems. Respecto a los estudios sobre dimensionalidad, permiten determinar la estructura del test, que puede haber sido construido para medir una o varias dimensiones, y ver si coincide con la estructura postulada al construir la prueba. Este tipo de análisis es frecuentemente realizado en los trabajos de validación. Se basa en el examen de las relaciones entre los ítems del test con el fin de determinar, empíricamente, qué conceptos se pueden emplear para interpretar sus puntuaciones. Se utilizan complejas técnicas estadísticas, fundamentalmente el análisis factorial, que examinan si las relaciones entre los ítems se corresponden con las hipotetizadas para el constructo que estamos midiendo. Por ejemplo, una teoría que plantea la unidimensionalidad de un constructo requiere que los ítems saturen en un único factor.

Mediante las técnicas factoriales, a partir de las correlaciones entre los ítems se obtiene una matriz factorial que expresa la relación entre los ítems y los factores comunes o dimensiones subyacentes. Los factores se definen como combinaciones lineales de los ítems originales.

El estudio de la dimensionalidad puede hacerse mediante diversos modelos de análisis factorial exploratorio (AFE) o análisis factorial confirmatorio (AFC). El AFE es básicamente una técnica de reducción de la dimensionalidad que permite pasar de un conjunto de variables observadas (ítems) a un número mucho menor de variables latentes o factores. El AFE busca identificar un conjunto de factores hipotéticos que pueden explicar las correlaciones observadas entre los ítems del test. No plantea hipótesis previas sobre las dimensiones y las saturaciones de los ítems en los factores. Los factores derivados del análisis son abstracciones matemáticas. Su significado sustantivo se desarrolla examinando el contenido de los ítems que saturan en cada factor. Por ejemplo, si todos los ítems que saturan en un factor implican habilidades de cálculo y los ítems que no requieren éstas habilidades tienen saturaciones muy bajas en él, el factor puede ser identificado como "Habilidad de cálculo". La interpretación surge al combinar el modelo matemático formal con juicios subjetivos que unen el modelo a fenómenos observables. En el capítulo 6 se describen con detalle el AFE.

El AFC, al igual que el exploratorio, tiene por objetivo identificar factores latentes que expliquen la covariación entre las variables observables. Ambos, AFE y AFC, están basados en el mismo modelo estadístico. La diferencia es que con el AFC se pone a prueba si una solución factorial concreta es o no adecuada para unos datos. Se especifica, por ejemplo, el número de factores, si están o no relacionados, qué ítems son indicadores de cada factor, etc. El AFC requiere una base empírica o conceptual fuerte que guíe la especificación del modelo. De ahí que se use en las últimas fases de los estudios de validación. El capítulo 10 recoge los aspectos técnicos más importantes para su aplicación.

Dentro de las evidencias relativas a la estructura interna también pueden ubicarse los trabajos encaminados a evaluar el funcionamiento diferencial de los ítems (FDI). El FDI aparece cuando personas con el mismo nivel en la característica medida por el test, pero que pertenecen a grupos distintos, tienen distinta probabilidad de acertar o estar de acuerdo con el ítem. Los grupos se definen atendiendo a variables sociodemográficas como el

sexo, la raza, cultura, idioma, etc. Una diferencia grupal no implica la existencia de FDI. Para hablar de FDI la diferencia entre los distintos grupos tiene que ser debida a diferencias en variables que no son las que el test pretende medir. El estudio del FDI también aporta evidencias sobre las consecuencias sociales del uso del test. Existen numerosas técnicas para detectar FDI, algunas de las cuales se describen en el capítulo 13.

Evidencias basadas en la relación con otras variables

El objetivo aquí es establecer si las relaciones observadas entre las puntuaciones en el test y otras variables externas relevantes son consistentes con la interpretación propuesta para las puntuaciones. Por ejemplo, Moltó (1988) predice (y comprueba) que la escala de susceptibilidad al castigo (que mide el grado de evitación de situaciones aversivas) debe proporcionar puntuaciones relacionadas directamente con neuroticismo e inversamente con estabilidad emocional. Si las relaciones observadas son consonantes con lo predicho por el modelo teórico en el que se inserta el constructo medido por el test, entonces hemos obtenido evidencia favorable a la interpretación propuesta. Si las relaciones observadas no son las esperadas hay que cuestionar no sólo la adecuación de la prueba, sino también la adecuación de las medidas de las otras variables e incluso la del modelo teórico.

Las variables externas relevantes a las que hacemos alusión pueden ser: a) otras medidas del mismo constructo obtenidas con diferentes tests; b) medidas de constructos diferentes pero que se insertan en el modelo teórico donde se encuadra el constructo de interés, o c) algún tipo de variable (criterio) que pretendamos predecir a partir de las puntuaciones en el test. Examinaremos por separado las evidencias para establecer la relación del test con otros constructos (evidencia convergente y discriminante) y del test con algún criterio (validez referida a un criterio).

La evidencia convergente y discriminante

Buscamos examinar las relaciones previsibles entre las puntuaciones en el test y otros constructos, ya sean similares (evidencia convergente) o diferentes (evidencia discriminante) al que se pretende medir con el test. Por ejemplo, podemos pensar que las puntuaciones en un test de opción múltiple de razonamiento lógico se relacionarán estrechamente con otra medida de razonamiento lógico basada en la resolución de problemas (evidencia convergente). Sin embargo, si medimos además otro constructo diferente, por ejemplo la comprensión lectora, esperamos que la relación entre ambas mediciones sea menor (evidencia discriminate). Predominan los trabajos que buscan obtener evidencia convergente, probablemente porque estudiar la relación entre distintos métodos que miden el mismo constructo puede ayudar a interpretar el significado de las puntuaciones.

Para obtener información sobre las relaciones entre las puntuaciones del test con otras variables que forman parte del modelo teórico se plantean habitualmente dos tipos de trabajos:

1. Estudios de comparación del rendimiento de diversos grupos en el test. Por ejemplo, en un test neuropsicológico podemos comparar grupos de personas con y sin lesión cere-

bral, o en un test de conocimientos un grupo de expertos con uno de novatos. En otras ocasiones se comparan grupos que han recibido intervenciones diferentes que deberían afectar a sus puntuaciones. Por ejemplo, puede estudiarse si las puntuaciones en una medida de stress son sensibles al tratamiento o si hay diferencias en las puntuaciones en un test de logro académico entre el grupo de estudiantes que han recibido instrucción y el que no la ha recibido. En otras ocasiones se comparan grupos para obtener evidencia discriminante. Por ejemplo, en un cuestionario que mida depresión (puntuaciones mal altas indicarían más depresión) podríamos comparar las puntuaciones obtenidas en el test por dos grupos de pacientes: un grupo con patología depresiva y otro grupo formado por pacientes con otros tipos de patologías. Si encontramos puntuaciones significativamente mas altas en el grupo de sujetos diagnosticados con depresión habríamos obtenido evidencias sobre un uso concreto del cuestionario.

Ejemplo 5.4. Un estudio sobre la validez convergente del listado de Psicopatía de Hare Revisado (PCL-R, Hare, 1991)

Chico y Tous (2003) estudiaron la validez convergente del listado de Psicopatía de Hare Revisado (PCL-R, Hare, 1991). En las últimas décadas el PCL-R se considera como el instrumento estandarizado más habitual para medir Psicopatía. Se aplicó la versión española de Moltó, Poy y Torrubia (2000) a una nuestra de 305 internos presos en un centro penitenciario. La escala consta de 20 ítems cuya puntuación viene determinada por el psicólogo, quien, usando la información obtenida en una entrevista semiestructurada, puntúa cada ítem como 0 (si la conducta en cuestión estaba ausente), 1 (si había dudas) o 2 (si se estaba seguro de su presencia). Para evaluar la validez convergente se observaba si existían diferencias estadísticamente significativas en variables relacionadas con la vida penitenciaria del recluso atendiendo a sus puntuaciones (altas o bajas) en el PCL-R. Se formaron dos grupos: el grupo 1 compuesto por presos que tenían puntuaciones por debajo de la media en la escala, y el grupo 2, formado por reclusos con puntuaciones por encima de la media. La Tabla 5.4 muestra que hubo diferencias en dos variables dependientes, en función de las puntuaciones altas y bajas en el PCL-R. Los presos que habían puntuado alto en PCL-R puntuaron más alto en la variable dependiente "número de ingresos en prisión" y eran más jóvenes cuando ingresaron por primera vez.

Tabla 5.4. Diferencias grupales en función de las puntuaciones en el PCL-R

Variables	Puntuaciones bajas en PCL-R ($N = 157$)		Puntuaciones altas en PCL-R ($N = 157$)		gl.	T
	Media	Desv. Tip.	Media	Desv. Tip.		
Edad 1er ingreso	21,14	3,70	18,50	3,45	323	6,40*
Nº de ingresos	3,09	2,66	5,45	4,54	323	−5,52*

*p < 0,0001.

También se había evaluado la gravedad de los delitos y la conducta en prisión. Los presos se clasificaron en función del delito mas grave que habían cometido en 3 categorías: 0, si habían cometido delitos no violentos; 1, si sus delitos suponían un cierto grado de violen-

cia (p. ej.: robos con fuerza); y 2, delitos mas violentos (p. ej.: robos con armas, violaciones, homicidios…). Respecto a su conducta, los reclusos se clasificaron de la siguiente forma: 0 (no tenían sanciones disciplinarias), 1 (tenían sanciones leves y como máximo una sola grave) y 2 (presos que habían cometido mas de una sanción grave o muy grave). Ambas variables estaban claramente relacionadas con la puntuación en el PCL-R: χ^2 (2, N = 305) = 89,56, $p < 0{,}001$ y χ^2 (2, N = 305) = 61,38, $p < 0{,}001$, respectivamente, para la tipología delictiva y la conducta en prisión.

2. En un segundo tipo de trabajos se obtienen las correlaciones entre las puntuaciones obtenidas en dos o más tests, para establecer si miden o no el mismo constructo. Si la previsión es que miden el mismo constructo, se estaría buscando una evidencia de validez convergente. Si la hipótesis de partida es que los tests miden constructos diferentes, se estaría buscando evidencia discriminante. Por ejemplo, Manners y Durkin (2001) realizan una revisión sobre las investigaciones realizadas para recoger evidencias sobre la validez del Washington University Sentence Completion Test (WUSCT), una escala desarrollada para medir el desarrollo del ego, construida desde la teoría de Loevinger sobre el desarrollo de la personalidad. Como ejemplos de trabajos que ofrecen evidencia discriminante para el WUSC, se citan varias investigaciones en las que se correlacionaron sus puntuaciones con medidas de fluidez verbal. Aunque ambos constructos, según predice la teoría, deben estar relacionados, ya que son necesarias más palabras para expresar mayor complejidad conceptual, el número de palabras usadas y la complejidad de las estructuras empleadas son claramente distinguibles, por lo que se esperaban obtener correlaciones medias. Este fue el resultado obtenido en distintas muestras, donde los coeficientes de correlación fueron aproximadamente de 0,30.

Ejemplo 5.5. Un estudio sobre la validez convergente de dos medidas objetivas de Minuciosidad

Hernández, Lozano, Shih y Santacreu (2009) realizaron una investigación para obtener un indicador de la validez convergente de dos medidas objetivas de minuciosidad que eran funcionalmente equivalentes; es decir, evaluaban el mismo estilo interactivo, que básicamente consiste en la ejecución de una tarea de manera ordenada, organizada, siguiendo un patrón sistemático. Este estilo podría considerarse equiparable, en términos teóricos, a la dimensión de Minuciosidad del Modelo de Cinco Factores de la Personalidad.

Las pruebas aplicadas fueron el Test de minuciosidad Árboles (TM-A) y el Test de Minuciosidad Fichas (TM-F). Hay apreciables diferencias formales entre ellos. La tarea en el test TM-A consistía en identificar y pulsar con el ratón, de entre una variedad de imágenes distintas, aquellas que fuesen iguales a una presentada como modelo. En el TM-F se presentaba un panel que contenía varios tipos de fichas, que otorgaban puntos al ser pulsadas. El objetivo de la tarea era obtener la mayor cantidad de puntos pulsando sobre el menor número de fichas.

Ambas pruebas de evaluación se aplicaron durante un proceso de selección. La distancia temporal entre una y otra fue de 1 hora y 40 minutos, período de tiempo en el cual los candidatos realizaron otras tareas de evaluación. Los candidatos ejecutaron las pruebas de modo individual en un ordenador. El coeficiente de correlación de Pearson entre las puntuaciones de las dos pruebas fue de 0,638, $p<0,001$. Los autores concluyen que ambas pruebas miden un mismo estilo interactivo. En otras palabras, las estrategias de actuación que ponen en marcha los individuos enfrentados a dos tareas distintas son las mismas. Estos resultados no permiten, no obstante, asegurar que estas pruebas estén midiendo la misma dimensión de minuciosidad que las tradicionales pruebas de evaluación basadas en el autoinforme de los individuos, ya que en un trabajo previo realizado por Sánchez-Balmisa, Hernández, Madrid, Peña y Santacreu (2003) no se encontró una correlación significativa entre el TM-F y la escala de responsabilidad del cuestionario de personalidad BFQ.

En 1959 Campbell y Fiske propusieron un diseño para analizar la validez convergente y discriminante, basado en el estudio de la denominada como *matriz multirrasgo-multimétodo*[3] (MRMM). Este trabajo es uno de los más citados en la historia de la Psicología. Para estos autores un test es el resultado de unir un constructo con un procedimiento de medida. Cuando las puntuaciones de dos instrumentos covarían puede deberse a que comparten un constructo común o a que comparten un método de evaluación. Para separar ambos aspectos, y así estudiar las contribuciones relativas de la varianza del constructo y del método, propusieron un diseño en el que una muestra de sujetos es evaluada en un conjunto de constructos, medidos cada uno con un conjunto de métodos diferentes. La matriz MRMM incluye todas las correlaciones entre condiciones de medida. El objetivo de estudiar una matriz MRMM es evaluar los efectos de la varianza atribuida al constructo de interés y la varianza del método (varianza atribuible al método de medida específico) ya que el efecto del método altera las correlaciones entre los constructos introduciendo sesgos sistemáticos. Idealmente, una medida no debería contener efecto del método. En contraste, los estudios MRMM han mostrado que las puntuaciones en los tests psicológicos y educativos contienen una cantidad sustancial de efecto del método (Dumenci, 2003). Adicionalmente, una MRMM también proporciona información sobre el patrón de asociaciones entre constructos y las posibles interacciones entre métodos y constructos.

Organización de las matrices MRMM

La selección de rasgos y métodos debe hacerse de modo que: a) cada uno de los métodos sea adecuado para medir todos los constructos de interés, b) los diferentes métodos sean lo más independientes entre sí y c) los constructos incluidos varíen en el grado de asociación entre ellos, con constructos altamente relacionados y otros en los que la asociación sea

[3] Para conmemorar el 50 aniversario de este trabajo la revista Methodology publicó en el año 2009 un número monográfico, Vol. 5 (3), en el que, para analizar los datos de esta matriz, presenta aproximaciones desde los modelos multinivel y el análisis factorial. El análisis de esta matriz aplicando el AFC puede consultarse en el capítulo 14.

muy baja. El objetivo de estas recomendaciones es establecer las condiciones para que las correlaciones entre las puntuaciones de diferentes rasgos, medidos con distintos métodos, se aproximen a cero.

La matriz MRMM se organiza por método, de modo que cada constructo medido está incrustado en cada bloque de método. Un ejemplo hipotético de organización lo podemos encontrar en la Tabla 5.5, que muestra una matriz de correlaciones para 3 constructos medidos por 2 métodos diferentes. Para interpretar esta matriz hay que identificar 4 regiones o grupos de correlaciones:

1. El primer grupo está formado por las correlaciones obtenidas entre los mismos constructos usando los mismos métodos (datos entre paréntesis). Son las correlaciones monorrasgo-monométodo y conforman las *diagonales de la fiabilidad.*
2. El segundo grupo lo forman las correlaciones entre las medidas del mismo constructo cuando se utilizan distintos métodos (datos en cursiva negrita). Son las correlaciones monorrasgo-heterométodo. Muestran evidencia sobre la convergencia y constituyen las *diagonales de la validez.*
3. El tercer grupo lo componen las correlaciones entre distintos constructos medidos con el mismo método (datos subrayados) o correlaciones heterorrasgo-monométodo. Nótese que estas correlaciones forman triángulos situados de forma adyacente a cada diagonal de la fiabilidad.
4. El cuarto grupo está formado por las correlaciones entre distintos constructos y distintos métodos, correlaciones en las que no se comparte ni el constructo ni el método, es decir, heterorrasgo-heterométodo. Nótese que forman triángulos adyacentes a la diagonal de la validez y que ambos triángulos no son iguales.

Ejemplo 5.6. Matrix MRMM para tres constructos medidos por dos métodos

En la Tabla 5.5 se presenta un ejemplo hipotético, adaptado de Fabrigar y Estrada (2007). Los métodos 1 y 2 son dos formas distintas de medir las actitudes. Los constructos *A*, *B* y *C* son, respectivamente, sentimientos, creencias e intenciones de actuar.

Tabla 5.5. Representación de una hipotética matriz multirrasgo-multimétodo

		Método 1			*Método* 2		
		Cons. A	*Cons. B*	*Cons. C*	*Cons. A*	*Cons. B*	*Cons. C*
Método 1	*Constructo A*	(0,98)					
	Constructo B	0,62	(0,95)				
	Constructo C	0,19	0,17	(0,93)			
Método 2	*Constructo A*	***0,75***	0,60	0,18	(0,95)		
	Constructo B	0,59	***0,86***	0,17	0,60	(0,94)	
	Constructo C	0,19	0,18	***0,74***	0,21	0,20	(0,95)

Interpretación de las matrices MRMM

El análisis tradicional de estas matrices, tal y como fue propuesto inicialmente por Campbell y Fiske, implica una inspección visual de la matriz examinando cuatro propiedades:

1. En primer lugar, hay que evaluar la diagonal monorrasgo-monométodo (o de la fiabilidad). Estos coeficientes deberían ser, de modo consistente, los más altos de la matriz, porque es poco probable que una medida correlacione más con cualquier otra cosa que consigo misma (por ejemplo, en dos aplicaciones). En nuestro ejemplo, las correlaciones varían entre 0,93 y 0,98 indicando valores elevados de la fiabilidad.
2. En segundo lugar, las correlaciones monorrasgo-heterométodo son tomadas como indicadores de evidencia convergente, porque nos informan del grado en que diferentes métodos son congruentes al medir el mismo constructo. Estas correlaciones deberían ser significativamente distintas de cero y lo suficientemente altas para que tenga sentido continuar un análisis de la validez. Idealmente, todos los métodos deberían proporcionar el mismo ordenamiento de los individuos para un particular constructo. En nuestro ejemplo, estas correlaciones son altas (varían entre 0,74 y 0,86) lo que sugiere que los diferentes métodos producen resultados similares para los tres constructos. El hecho de que estas correlaciones sean elevadas es una condición necesaria, pero no suficiente, para asegurar la convergencia. Es posible que estas correlaciones estén sobreestimadas por un factor irrelevante (por ejemplo, la varianza del método), y por eso es necesario examinar las correlaciones que nos proporcionan evidencia sobre la divergencia, tal como se indica a continuación.
3. En tercer lugar, las correlaciones monorrasgo-heterométodo hay que compararlas con los triángulos heterorrasgo-monométodo. Los valores en la diagonal monorrasgo-heterométodo deberían ser mas altos que los valores de los triángulos heterorrasgo-monometodo, porque distintos métodos evaluando un mismo rasgo deberían correlacionar más que el mismo método evaluando rasgos distintos. Si no ocurriese esto, el método de medida explicaría una parte importante de varianza de las puntuaciones. En nuestro ejemplo, las correlaciones monorrasgo-heterométodo (0,75, 0,86, y 0,74) son, para cada comparación, mayores que las correlaciones obtenidas en los triángulos heterorrasgo-monométodo: 0,62, 0,19 y 0,17 (para el Método 1) y 0,60, 0,21 y 0,20 (para el Método 2). Se debe cumplir también que las correlaciones monorrasgo-heterométodo sean mas altas que las obtenidas en los triángulos heterorrasgo-heterométodo para la misma fila o columna. Esencialmente, si diferentes métodos están midiendo el mismo constructo, sus correlaciones deberían ser mayores que las de constructos distintos que están medidos usando métodos distintos. Por ejemplo, en nuestra matriz 0,75 es mayor que las correlaciones de su fila (0,60 y 0,18); y también es mayor que las correlaciones de su columna (0,59 y 0,19). La misma propiedad se cumple para 0,86 (mayor que 0,59, 0,17, 0,60 y 0,18) y también para 0,74 (mayor que 0,19, 0,18, 0,18 y 0,17).
4. En cuarto lugar, y para terminar, el investigador debe comparar los triángulos heterorrasgo-monométodo y heterorrasgo-heterométodo, ya que si dos rasgos están correlacionados, esta relación debería mantenerse con independencia del método utilizado para medirlos y el mismo patrón debería estar visible en todos los bloques monométodo y heterométodo. Si examinamos los datos de nuestro ejemplo, este criterio se satisface siempre. Además, para aquellos constructos que estén correlacionados, las correlacio-

nes heterorrasgo-heterométodo deberían ser mas altas que para los constructos que no lo estén. La inspección visual de la Tabla 5.5 nos indica que la regla anterior se cumple para cada comparación. Veámoslo, si medimos los constructos con el mismo método, encontramos que la relación entre los constructos A y B es más alta (0,62 y 0,60) que la existente entre los constructos A y C (0,19, 0,21) y también que la obtenida entre los constructos B y C (0,18, 0,20). Al comparar las relaciones entre los constructos cuando son medidos con distintos métodos, la relación entre los constructos A y B (0,59 y 0,60) sigue siendo mayor que la obtenida entre los contructos A y C (0,19, 0,18) y que la obtenida entre los contructos B y C (0,17, 0,18). También se obtiene evidencia sobre el efecto del método al examinar la magnitud diferencial de las correlaciones entre dos constructos diferentes medidos por el mismo método y las correlaciones entre los mismos dos constructos medidos por distintos métodos. Por ejemplo, los constructos A y B correlacionan 0,62 ó 0,60, según se midan con el Método 1 o con el Método 2. Cuando se miden con métodos distintos las correlaciones difieren muy poco (0,59 y 0,60).

En resumen, una matriz MRMM debería proporcionar evidencia a favor de la convergencia al medir los mismos constructos, de la divergencia al medir constructos distintos, y de la ausencia de efectos del método. El estudio de este tipo de matrices tiene también algunas limitaciones. Por una parte, algunas asunciones claves subyacentes no están claramente definidas. Por ejemplo, en el estudio de la matriz MRMM se asume la existencia de dos tipos de variables (método y rasgo), pero no se especifica su interacción; tampoco se considera el efecto del error de medida en la cuantía de las correlaciones. Por otra parte, hay algunos problemas prácticos asociados con su uso. Por ejemplo, no siempre es posible disponer de un diseño completo método x rasgos; además, las matrices muy grandes pueden ser muy complicadas de evaluar. También se ha criticado la ambigüedad de la interpretación, dado que a veces se producen resultados contradictorios dentro de una misma matriz. Así, en la práctica es habitual que algunos aspectos de la matriz sean consistentes con las reglas de interpretación, mientras que otros pueden no serlo. En tales casos las evaluaciones de las diferentes correlaciones pueden ser muy subjetivas. Con el objetivo de afrontar la dificultad de interpretación de las matrices MRMM, y así cuantificar el grado en que tales criterios han sido satisfechos, se han desarrollado diversos procedimientos estadísticos para complementar la evaluación visual. En el capítulo 14 se expone con algún detalle el estadístico recientemente propuesto por Sawilowsky (2002) y el empleo del AFC para complementar y aclarar la interpretación.

Evidencia sobre la relación entre el test y algún criterio relevante

Cuando se pretende utilizar el test para pronosticar determinados criterios de rendimiento como, por ejemplo, el rendimiento escolar, el total de ventas que se van a conseguir, el aprovechamiento conseguido en un cursillo o la mejora en un proceso terapéutico, se requiere que el test se relacione muy estrechamente con dichos criterios. Suele hablarse entonces de la necesidad de obtener *evidencias de validez referida al criterio*, lo cual requiere:

1. Identificar un criterio y la manera adecuada de medirlo.
2. Elegir una muestra apropiada.
3. Obtener en la muestra medidas en el test y en el criterio.

4. Determinar el grado de relación ente ambos.

Para obtener la relación entre el test (X) y el criterio (Y), si son variables continuas, se calcula la correlación entre ambas variables, que se denomina *coeficiente de validez* (r_{XY}) e indica el grado en que las puntuaciones en el test sirven para pronosticar con precisión las puntuaciones en el criterio. Supongamos, por ejemplo, que la correlación entre un test de conocimientos y las calificaciones obtenidas en 2º de Bachillerato fuese 0,85 en una muestra representativa. Como la correlación es elevada, cometeríamos errores de pronóstico no excesivamente elevados (haciendo uso de la oportuna ecuación de regresión) al predecir la calificación de un alumno sabiendo su rendimiento en el test de conocimientos. El coeficiente de validez no es una propiedad del test, sino que habrá un coeficiente específico en cada muestra donde se obtenga y para los diferentes criterios que puedan establecerse.

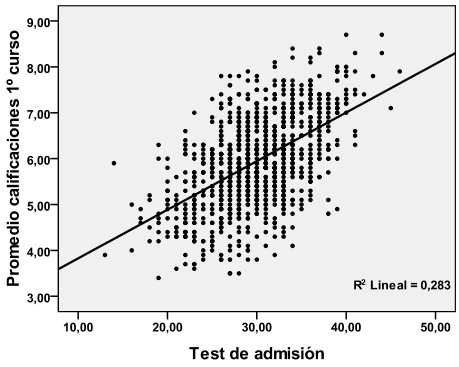
Cuando las puntuaciones en los tests van a emplearse para tomar decisiones importantes para los evaluados (p.ej., sobre su admisión o no a un puesto de trabajo, sobre el acceso a una plaza escolar determinada o sobre su acreditación profesional) es preciso que los profesionales dispongan de tests con elevada validez referida al criterio. Pero esto resulta a veces difícil o muy laborioso. En muchas ocasiones no resulta sencillo establecer criterios pertinentes (relacionados con el test), fiables y fácilmente mensurables, lo cual afectará a la precisión con la que podrán estimarse. Por ejemplo, los tests que se emplean en el examen teórico para obtener el permiso de conducir, deberían predecir en parte la habilidad futura de conducción, un criterio que probablemente no resulta sencillo de medir de forma fiable; además, seguramente resultaría muy costoso obtener evidencias de validez referida a este criterio para todos los diferentes tests teóricos que se aplican.

La validez referida a un criterio puede ser *predictiva* o *concurrente*. La distinción entre ambas se refiere al intervalo de tiempo transcurrido entre las mediciones en el test y en el criterio. Las evidencias de validez predictiva reflejan la relación entre las puntuaciones en un test y un criterio, cuando el criterio se mide más tarde. Por ejemplo, si en un proceso de selección de personal se aplica un test de aptitudes cognitivas, podrá correlacionarse con medidas de desempeño laboral sólo después de que los admitidos tengan la oportunidad de trabajar durante un tiempo. En el caso de la validez concurrente, las medidas en el test y en el criterio se obtienen aproximadamente en el mismo momento.

Interpretación del coeficiente de validez

Si las puntuaciones en el test (X) y en el criterio que se desea pronosticar (Y) son variables continuas, el modelo de regresión lineal simple permite cuantificar la capacidad predictiva del test. La hipótesis básica del modelo es la linealidad de la relación entre ambos. La función que relaciona las puntuaciones en el test con las del criterio deberá tener un incremento (o decremento) constante para los diferentes valores de X. Un diagrama de dispersión, como el que se representa en la Figura 5.1, nos permite obtener una aproximación sencilla al estudio del grado de relación lineal. Es importante complementar el cálculo del coeficiente de validez con el correspondiente diagrama de dispersión, ya que un mismo coeficiente puede ser obtenido con distintas pautas de relación y el diagrama es una forma sencilla de visualizar estas pautas. En la Figura 5.1 se recogen los datos, obtenidos por simulación, de una muestra de 1.000 estudiantes en un hipotético test de admisión al centro (X) y el promedio de sus calificaciones obtenidas en el primer curso del grado en Psicología (Y).

Figura 5.1. Diagrama de dispersión de *Y* (calificación) sobre *X* (puntuaciones en un test de admisión). Se ha simulado una muestra de 1.000 estudiantes



Los alumnos con puntuaciones más altas (bajas) en el examen de admisión tienden a obtener una calificación promedio más elevada (baja) durante el primer curso del grado. En nuestro ejemplo la correlación entre ambas variables (coeficiente de validez) fue de 0,532, que indica una relación lineal positiva entre el test y el criterio[4].

La recta de regresión que se ha dibujado es la línea que mejor se ajusta a la nube de puntos y nos permite predecir la calificación que obtendría un estudiante que haya obtenido una puntuación concreta en el test. La distancia vertical entre un punto y la línea de regresión es el error de pronóstico o residuo para ese punto. La recta de regresión se ha calculado usando el método de estimación más habitual, mínimos cuadrados ordinarios, que minimiza la suma de los errores al cuadrado. En nuestro caso, la capacidad predictiva del test no es muy elevada, ya que la mayor parte de los puntos distan bastante de la recta.

El coeficiente de validez es una correlación de Pearson y, por tanto, su interpretación más inmediata se fundamenta en el *coeficiente de determinación*, que es el cuadrado de la correlación y que indica la proporción de varianza que comparten las puntuaciones del test y del criterio. Así, el coeficiente de validez de 0,532 de nuestro ejemplo indica que con el test se explica un 28,3 % de la variabilidad o diferencias individuales en el criterio, mientras que el 71,7 % restante se debe a variables diferentes al test (errores de medida en ambos y otras variables no contempladas que influyen en las calificaciones). Recordando algunos conceptos fundamentales de la regresión lineal simple, el coeficiente de determinación se puede expresar como:

---

[4] Como veremos un poco más adelante, puede obtenerse también la significación estadística de la correlación (contrastar si es diferente de 0 en la población). En este sentido, conviene recordar la incidencia del tamaño de la muestra, de modo que puede alcanzarse la significación para coeficientes bajos cuando están obtenidos en muestras de gran tamaño. Una correlación significativa puede no ser una correlación elevada. Generalmente los coeficientes de validez no exceden de 0,6 en situaciones reales.

$$r_{XY}^2 = \frac{S_{Y'}^2}{S_Y^2} = 1 - \frac{S_{Y-Y'}^2}{S_Y^2} \qquad [5.2]$$

Donde:

S_Y^2 es la varianza del criterio,

$S_{Y'}^2$ es la varianza de los pronósticos,

$S_{Y-Y'}^2$ es la varianza de los errores de pronóstico.

Si conocemos el coeficiente de validez y la varianza de las puntuaciones del criterio, podremos obtener la varianza de los errores de pronóstico despejando de la ecuación [5.2]:

$$S_{Y-Y'}^2 = S_Y^2 \sqrt{1 - r_{XY}^2} \qquad [5.3]$$

La desviación típica de los errores de pronóstico ($S_{Y\text{-}Y'}$) recibe el nombre de *error típico de estimación* y juega un importante papel en las aplicaciones.

Estimaciones en el criterio

La función lineal que permite predecir las puntuaciones en el criterio a partir de las puntuaciones en el test será:

$$Y_i' = \beta_0 + \beta_1 X_i \qquad [5.4]$$

Donde β_0 es la constante, ordenada en el origen o intercepto y representa el valor esperado de Y cuando X toma el valor 0, y β_1 es la pendiente de la recta o coeficiente de regresión (muestra el cambio que experimenta el valor de Y cuando X cambia una unidad). Gráficamente, β_0 representa el punto en el que la recta de regresión corta el eje de ordenadas y β_1 representa la iclinación de la recta. Como la relación entre X e Y no es exacta, para cada sujeto i cometemos algún error de pronóstico ($Y_i - Y_i'$). Cuanto más próximo esté un punto a la recta de regresión, menor será el error cometido.

Para determinar los valores de β_0 y de β_1 puede utilizarse el criterio denominado de *mínimos cuadrados ordinarios* que minimiza la suma de los errores al cuadrado para el conjunto de los sujetos:

$$\sum_{i=1}^{N} \left(Y_i - Y_i'\right)^2 \qquad [5.5]$$

La recta que hace mínima la expresión [5.5] se consigue sustituyendo $Y_i^{'}$ por su valor $Y_i^{'} = \beta_0 + \beta_1 X_i$. El proceso de minimización conduce a dos ecuaciones de las que se pueden despejar los valores de los dos parámetros. Puesto que se trabaja con datos muestrales:

$$b_1 = r_{XY} \frac{S_Y}{S_X} \qquad [5.6]$$

$$b_0 = \overline{Y} - b_1 \overline{X} \qquad [5.7]$$

En nuestro ejemplo, las desviaciones típicas del criterio y del test son, respectivamente, 0,973 y 4,886; las correspondientes medias son 5,927 y 29,818. Si quisiéramos predecir la calificación promedio en el primer curso a partir de las puntuaciones en el examen de admisión, la ecuación de regresión se obtendría de la siguiente forma:

$$b_1 = r_{XY} \frac{S_Y}{S_X} = 0{,}532 \frac{0{,}973}{4{,}886} = 0{,}106$$

$$b_0 = \overline{Y} - b_1 \overline{X} = 5{,}927 - 0{,}106(29{,}818) = 2{,}768$$

$$Y_i^{'} = b_0 + b_1 X_i = 2{,}768 + 0{,}106 X_i$$

Por lo tanto, en la ecuación anterior b_1 nos indica que un incremento de un punto en la nota del examen de admisión, produce un aumento de 0,106 puntos en la calificación promedio en el primer curso.

El valor obtenido para el estimador de la pendiente variará al calcularlo en distintas muestras, aunque procedan de la misma población. Estos valores constituyen la distribución muestral del coeficiente de regresión; el tamaño de la variación vendrá indicado por el error típico de estimación de dicho parámetro, en inglés Standard Error (SE):

$$SE_{b_1} = \frac{S_Y}{S_X} \sqrt{\frac{1 - r_{XY}^2}{N - 2}} \qquad [5.8]$$

Donde N es el tamaño de la muestra. El intervalo de confianza para el coeficiente de regresión se obtiene mediante la expresión:

$$b_1 \pm t_{1-\alpha/2} SE_{b_1} \qquad [5.9]$$

Donde t es el valor de la distribución t de Student con $N-2$ grados de libertad y probabilidad $1-\alpha/2$. Si este intervalo incluyese el valor de cero, entonces no podríamos rechazar la hipótesis nula de que, en la población, el coeficiente de regresión sea cero.[5]

En nuestro ejemplo, el error típico de estimación del coeficiente de regresión es:

$$SE_{b_1} = \frac{0{,}973}{4{,}886}\sqrt{\frac{1-0{,}283}{1.000-2}} = 0{,}005$$

Y el intervalo, con un nivel de confianza del 95%, será:

$$0{,}106 \pm ({}_{998}t_{0{,}975})0{,}005 = 0{,}106 \pm 1{,}960(0{,}005) = 0{,}106 \pm 0{,}0098$$

Por lo que en la población el valor del coeficiente de regresión estará comprendido entre 0,096 y 0,116 con una probabilidad del 95%. Dicho de otro modo, el coeficiente de validez ha resultado estadísticamente significativo, lo cual no garantiza que las estimaciones en el criterio ser realicen con precisión.

La ecuación de regresión de Y sobre X puede expresarse también (para puntuaciones directas) como:

$$Y_i' = (\overline{Y} - r_{XY}\frac{S_Y}{S_X}\overline{X}) + r_{XY}\frac{S_Y}{S_X}X_i \qquad [5.10]$$

Si queremos utilizar puntuaciones diferenciales, es decir, manteniendo la desviación típica original, pero con media cero en X e Y, la ecuación de regresión es:

$$y_i' = r_{XY}\frac{S_Y}{S_X}x_i \qquad [5.11]$$

Si deseamos usar puntuaciones típicas, donde las medias serán cero y las desviaciones típicas de X e Y serán uno, entonces la ecuación es:

$$Z_{Y_i}' = r_{XY}Z_{X_i} \qquad [5.12]$$

Como puede observarse, la pendiente en la ecuación de regresión para puntuaciones típicas, también denominado *coeficiente de regresión estandarizado* o *peso beta*, es el coeficiente de

---

[5] Si queremos aplicar la ecuación que hemos obtenido en nuestra muestra a otra muestra que proceda de la misma población, es decir para hacer un uso inferencial de nuestra ecuación, necesitaremos suponer que en la población se cumplen ciertas características o supuestos. Afortunadamente los estadísticos empleados en la regresión lineal simple son robustos, es decir, desviaciones moderadas de los supuestos no producen errores graves en la inferencia. Básicamente los supuestos hacen referencia a la distribución normal y a la homocedasticidad de los errores de predicción; la falta de homocedasticidad implicaría que los errores que cometiésemos para los distintos valores de X no serían de la misma magnitud; por ejemplo, a valores grandes de X le corresponderían valores grandes del error.

correlación de Pearson. Es donde mejor podemos ver que las estimaciones en *Y* serán tanto más precisas cuanto mayor sea r_{XY}.

Nótese que el valor de la ordenada en el origen de las ecuaciones en puntuaciones diferenciales y típicas es cero, por lo tanto, ambas rectas cruzarán el origen de coordenadas. La pendiente de la recta de regresión en puntuaciones directas y diferenciales es la misma, por lo que ambas rectas serán paralelas; pero la pendiente en puntuaciones típicas es por lo general distinta, y por tanto esta recta no será paralela a las anteriores.

La ecuación de regresión para puntuaciones típicas, correspondiente a los datos del ejemplo, se muestra a continuación; nos indica que por cada desviación típica de aumento en Z_X se produce un aumento de 0,532 desviaciones típicas en las puntuaciones típicas de calificación.

$$Z'_{Y_i} = (0{,}532) Z_{X\,i}$$

Hasta ahora se han realizado estimaciones puntuales en *Y*. Estadísticamente, resulta más apropiada una estimación por intervalos, realizada con cierta probabilidad, para lo cual aplicaremos la siguiente expresión:

$$Y'_i \pm Z_{1-\alpha/2} S_{Y-Y'} \qquad [5.13]$$

Donde $Z_{1-\alpha/2}$ es el valor de la distribución *N (0, 1)*, que deja por debajo la probabilidad $1 - \alpha/2$. y $S_{Y-Y'}$ es el error típico de estimación definido en la expresión [5.3].

Ejemplo 5.7. Intervalo de confianza para una puntuación pronosticada

A una muestra de 5 estudiantes de bachillerato se le aplica un test de habilidades comunicativas (*X*). A sus respectivos profesores se les pide que hagan una valoración (de 0 a 20 puntos) de la capacidad de relación interpersonal de sus alumnos. Estas valoraciones hacen la función de criterio (*Y*). Los resultados en el test y en el criterio se muestran en las columnas *X* e *Y* de la Tabla 5.6.

Tabla 5.6. Puntuaciones en un test de habilidades comunicativas y un criterio (capacidad de relación interpersonal) en una muestra de 5 estudiantes

Alumno	*X*	*Y*	*Y′*	$Y - Y'$
1	7	6	6,6	–0,6
2	13	10	11,4	–1,4
3	10	9	9	0
4	9	8	8,2	–0,2
5	11	12	9,8	2,2
Media	10	9		
S_X	2,236	2,236		

El coeficiente de validez del test es $r_{XY} = 0,8$, lo que significa que el test de habilidades comunicativas explica un 64 % de las diferencias en las valoraciones de los profesores sobre la capacidad de relación interpersonal de los estudiantes.

Para realizar una estimación puntual de la puntuación en el criterio de un estudiante, aplicamos la ecuación de regresión [5.10]. Los estimadores de los pesos de la ecuación de regresión serían:

$$b_1 = r_{XY}\frac{S_Y}{S_X} = 0,8\frac{2,236}{2,236} = 0,8$$

$$b_0 = \overline{Y} - b_1\overline{X} = 9 - 0,8(10) = 1$$

Y la ecuación de regresión[6]:

$$Y_i^{'} = b_0 + b_1 X_i = 1 + 0,8X_i$$

La tabla 5.6 recoge en las dos últimas columnas los pronósticos y los errores de pronóstico cometidos para cada estudiante. Por ejemplo, al nº 5 le pronosticamos una puntuación en el criterio $Y_5' = 9,8$ y cometemos un error de pronóstico de $Y_5 - Y_5^{'} = 12 - 9,8 = 2,2$ puntos.

Para realizar la estimación por intervalos para este mismo estudiante, con probabilidad 0,95, fijamos el valor $Z_{1-\alpha/2} = 1,96$ y calculamos el error típico de estimación:

$$S_{Y-Y'} = S_Y\sqrt{1 - r_{XY}^2} = 1,342$$

y el intervalo será:

$$Y_i^{'} \pm Z_{1-\alpha/2}S_{Y-Y'} = 9,8 \pm (1,96)(1,342) = 9,8 \pm 2,629$$

Diremos entonces que, con probabilidad 0,95, la puntuación de este estudiante en el criterio se encontrará comprendida entre 7,171 y 12,429. Como vemos, la amplitud del intervalo es amplia (algo no deseable) a pesar de que el coeficiente de validez era elevado.

Lo que ocurre en el ejemplo es ilustrativo de lo difícil que resulta realizar pronósticos precisos a partir de las puntuaciones en un único test. Cuando se desea predecir de la forma más precisa posible las puntuaciones en un criterio, es común utilizar las puntuaciones en varias varia-

[6] Nótese que, en este ejemplo, el coeficiente de correlación de Pearson y b_1 toman el mismo valor, ya que el test y el criterio tienen la misma varianza.

bles predictoras X (p.ej., en diferentes tests)[7]. En este caso, los pronósticos se realizarán con la técnica estadística de *Análisis de Regresión Múltiple*, que proporciona los pesos (*coeficientes de regresión parcial*) de cada predictor según la importancia que tengan para la predicción. Así, la ecuación de regresión múltiple será:

$$Y_i^{'} = b_0 + b_1 X_{1i} + \cdots + b_k X_{ki} + \cdots + b_K X_{Ki} \qquad [5.14]$$

Donde K es el número de variables predictoras.

Un tratamiento más amplio del modelo de regresión lineal aplicado a las Ciencias Sociales puede encontrarse, entre otros, en Cohen, Cohen, West, y Aiken (2003) y, en castellano, en los libros de Etxeberría (1999) y Pardo, Ruiz y San Martín (2009).

En el capítulo 14 se comentan con más detalle los distintos aspectos relacionados con la regresión lineal múltiple. Si el criterio a pronosticar fuese una variable discreta se pueden utilizar otras técnicas multivariadas, como el análisis discriminante y la regresión logística, tal como veremos también en ese capítulo. En algunos contextos aplicados es especialmente importante comprobar que la ecuación de regresión es la misma en diferentes submuestras (p.ej., de mujeres y hombres). Se trata de estudiar *la validez predictiva diferencial* (o evidencias externas de sesgo), tal como veremos en el capítulo 13.

Factores que afectan al coeficiente de validez

La cuantía de la correlación entre el test y el criterio (y por tanto la precisión de los pronósticos) viene condicionada por varios factores, entre los cuales están:

1. La fiabilidad del test.
2. La fiabilidad del criterio.
3. La auténtica relación entre test y criterio.
4. Características de la muestra, como es su tamaño, representatividad y su variabilidad en el test y en el criterio.

Además, es importante que los errores de medida en el test y en el criterio sean independientes. Es decir que el coeficiente de validez obtenido refleje la relación verdadera entre la dos variables, y no sea debida en parte a otras variables extrañas e irrelevantes, como puede ser un criterio contaminado. Esto puede suceder, por ejemplo, cuando la misma persona que administra el test (y conoce las puntuaciones obtenidas) valora subjetivamente a las personas en el criterio; su conocimiento de los resultados en el test puede sesgar sus valoraciones Y.

[7] Por ejemplo, en los estudios sobre predicción del desempeño laboral se ha encontrado que la mejor combinación de predictores es la formada por un test de capacidad cognitiva general, una entrevista conductual estructurada y un test de personalidad que mida el factor de Responsabilidad (Salgado y Moscoso, 2008).

Fiabilidad del test y del criterio

El coeficiente de validez depende del nivel de precisión con que se miden las puntuaciones en el test y en el criterio. Una baja fiabilidad, ya sea en *X* ó en *Y*, hará que el coeficiente de validez obtenido subestime la relación entre las puntuaciones verdaderas en el test y en el criterio. Si el coeficiente de fiabilidad de un test es bajo, existe una parte importante de error en las puntuaciones *X* que, al ser aleatorio, no contribuirá a la correlación entre *X* e *Y*; en ese caso, el coeficiente de validez obtenido será sensiblemente menor que la correlación entre las puntuaciones verdaderas en ambos. Lo mismo se puede decir para niveles altos de error de medida en el criterio.

Atenuación es el término que se usa para describir la reducción en la magnitud de la correlación entre dos medidas que está causada por su falta de fiabilidad. Spearman[8] fue el primero en reconocer el valor de corregir por atenuación, al señalar que estamos interesados en determinar la verdadera relación entre los constructos que estudiamos, y no tanto la relación entre dos medidas empíricas con más o menos error. Su solución fue estimar la correlación que habría entre las puntuaciones en el test y en el criterio si ambos fueran perfectamente fiables.

Bajo ciertos supuestos, puede comprobarse que el límite máximo al que puede llegar r_{XY} es $\sqrt{r_{XX} r_{YY}}$. Es decir, que:

$$r_{xy} \leq \sqrt{r_{XX} r_{YY}} \qquad [5.15]$$

Donde r_{XX} es el coeficiente de fiabilidad del test y r_{YY} es el coeficiente de fiabilidad del criterio. La desigualdad anterior indica que el coeficiente de validez viene determinado en parte por el coeficiente de fiabilidad del test y del criterio.

Veamos cómo se obtiene la relación expresada en la ecuación [5.15]. Una de las expresiones de la correlación de Pearson es:

$$r_{XY} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{N S_X S_Y} \qquad [5.16]$$

Si consideramos que los supuestos de la TCT se cumplen tanto en el test como en el criterio, pueden realizarse las sustituciones oportunas (recuerde que la media de los errores en el test y en el criterio es cero) para llegar a una expresión equivalente a [5.16]:

$$r_{XY} = \frac{\sum (V_X + E_X - \overline{V}_X)(V_Y + E_Y - \overline{V}_Y)}{N S_X S_Y} = \frac{\sum (v_X + e_X)(v_Y + e_Y)}{N S_X S_Y}$$

[8] Fan (2003) propone el AFC como una segunda manera de corregir por atenuación. En el AFC el error de medida de cada variable latente es explícitamente modelado. En ambos procedimientos se encuentran resultados altamente comparables para los mismos datos. Sin embargo, el AFC puede ser menos aplicable dadas las restricciones del modelo sobre los datos de los ítems (p. ej.: extrema asimetría y curtosis, distribuciones diferentes de los ítems, etc.).

Si en la segunda expresión se calculan los productos término a término en el numerador, divididos entre *N*, se obtienen covarianzas entre las diversas variables. Aplicando los supuestos cuarto y quinto del modelo clásico, que asumen una relación nula entre puntuaciones verdaderas y errores (y entre errores en diferentes tests), se anulan las covarianzas entre *v* y *e*, de tal forma que el coeficiente de validez vendría expresado como:

$$r_{XY} = \frac{Cov(V_X, V_Y)}{S_X S_Y} \qquad [5.17]$$

Otra manera de expresar la ecuación anterior es:

$$r_{XY} = \frac{r_{V_X V_Y} S_{V_X} S_{V_Y}}{S_X S_Y} = r_{V_X V_Y} \sqrt{r_{XX} r_{YY}} \qquad [5.18]$$

Dado que la correlación entre puntuaciones verdaderas en el test y puntuaciones verdaderas en el criterio es igual o inferior a 1, queda demostrada la desigualdad [5.15]. Imaginemos, por ejemplo, que un test de inteligencia general manifiesta un $r_{XX} = 0{,}85$, mientras que una prueba de cultura general, considerada como criterio, manifiesta un $r_{YY} = 0{,}73$. Según estos valores de los coeficientes de fiabilidad, el coeficiente de validez de este test respecto a este criterio no puede superar el valor de 0,79, que es la raíz cuadrada del producto entre los dos coeficientes de fiabilidad.

Ejemplo 5.8. Aplicación de la corrección por atenuación

Supongamos que un investigador desea conocer la validez de las puntuaciones en un test de "Afectividad negativa", entendida como la tendencia general a experimentar emociones negativas, para predecir las puntuaciones en una escala de "Satisfacción laboral". En una muestra de empleados, la correlación entre el test y el criterio fue 0,40. Dado que ambas puntuaciones están afectadas por errores de medida, aplica la corrección por atenuación utilizando los coeficientes de fiabilidad del test (0,84) y del criterio (0,76). Para ello, despejando la correlación entre puntuaciones verdaderas en la expresión [5.18].

$$r_{V_X V_Y} = \frac{r_{XY}}{\sqrt{r_{XX} r_{YY}}} = \frac{0{,}40}{\sqrt{0{,}84 \cdot 0{,}76}} = 0{,}50$$

El nuevo coeficiente de validez, estimado como la correlación entre puntuaciones verdaderas (y por tanto después de corregir la atenuación) toma el valor de 0,50.

De lo expuesto hasta ahora se deduce además que, dado que el valor máximo de un coeficiente de fiabilidad es uno, el coeficiente de validez de un test es menor o igual que la raíz cua-

drada del coeficiente de fiabilidad del test; también es menor o igual que la raíz cuadrada de la fiabilidad del criterio:

$$r_{XY} \leq \sqrt{r_{XX} r_{YY}} \leq \sqrt{r_{XX}} \qquad [5.19]$$

$$r_{XY} \leq \sqrt{r_{XX} r_{YY}} \leq \sqrt{r_{YY}} \qquad [5.20]$$

La ecuación general a partir de la cual se pueden estimar los cambios producidos en el coeficiente de validez cuando cambian los coeficientes de fiabilidad del test y del criterio (p.ej. porque se alargan con formas paralelas) es la siguiente:

$$r_{X_2Y_2} = \frac{r_{X_1Y_1}}{\sqrt{\frac{r_{X_1X_1} r_{Y_1Y_1}}{r_{X_2X_2} r_{Y_2Y_2}}}} \qquad [5.21]$$

Donde:

$r_{X_2Y_2}$ es el coeficiente de validez cuando se modifica la fiabilidad del test y la del criterio.

$r_{X_1Y_1}$ es el coeficiente de validez del test y criterio iniciales.

$r_{X_1X_1}$ es el coeficiente de fiabilidad del test inicial.

$r_{X_2X_2}$ es el coeficiente de fiabilidad del test modificado

$r_{Y_1Y_1}$ es el coeficiente de fiabilidad del criterio inicial

$r_{Y_2Y_2}$ es el coeficiente de fiabilidad del criterio modificado

La ecuación [5.21] se puede demostrar del modo siguiente. Según [5.18] los coeficientes de validez del test inicial y del test modificado serían, respectivamente:

$$r_{X_1Y_1} = r_{V_XV_Y} \sqrt{r_{X_1X_1} r_{Y_1Y_1}}$$

$$r_{X_2Y_2} = r_{V_XV_Y} \sqrt{r_{X_2X_2} r_{Y_2Y_2}}$$

Si despejamos $r_{V_XV_Y}$ en la primera expresión y sustituimos su valor en la segunda, tendremos que:

$$r_{X_2Y_2} = \left(\frac{r_{X_1Y_1}}{\sqrt{r_{X_1X_1} r_{Y_1Y_1}}} \right) \sqrt{r_{X_2X_2} r_{Y_2Y_2}} = \frac{r_{X_1Y_1}}{\sqrt{\frac{r_{X_1X_1} r_{Y_1Y_1}}{r_{X_2X_2} r_{Y_2Y_2}}}}$$

Si sólo modificásemos la fiabilidad del test, $r_{Y_1Y_1} = r_{Y_2Y_2}$, con lo que el valor del coeficiente de validez del test modificado respecto al criterio inicial sería:

$$r_{X_2Y_1} = \frac{r_{X_1Y_1}}{\sqrt{\frac{r_{X_1X_1}}{r_{X_2X_2}}}} \qquad [5.22]$$

De modo análogo, si sólo modificamos la fiabilidad del criterio, el valor del coeficiente de validez del test inicial respecto al criterio modificado sería:

$$r_{X_1Y_2} = \frac{r_{X_1Y_1}}{\sqrt{\frac{r_{Y_1Y_1}}{r_{Y_2Y_2}}}} \qquad [5.23]$$

Lo que se olvida a veces en los procesos de obtención de evidencias sobre la validez referida a un criterio es que el coeficiente de validez depende no sólo de la precisión de la medida que ofrece el test, sino también de la precisión con que medimos el criterio.

Otro asunto a considerar es el tipo de coeficiente de fiabilidad que debe ser usado para realizar la corrección por atenuación: test-retest, formas paralelas o consistencia interna. Los valores de unos y otros pueden diferir para una misma aplicación del test (y también para estimar la fiabilidad de las puntuaciones en el criterio). Por otra parte, sabemos que los diferentes métodos capturan diferentes componentes del error. Si se pierde un componente del error que es importante para la situación o contexto estudiado, entonces la corrección por atenuación puede no representar la correlación entre puntuaciones verdaderas. Lo más aconsejable es elegir uno u otro coeficiente en función de los objetivos pretendidos en el estudio de validez referida al criterio. Por ejemplo, si el interés es conocer la capacidad de un test para predecir, un año más tarde, el logro académico de los estudiantes, entonces deberíamos emplear un coeficiente de fiablidad test-retest. Si, por el contrario, estamos interesados en incrementar el número de ítems de un test con el objetivo de mejorar su capacidad pronóstica, entonces las estimaciones basadas en la consistencia interna (p.ej., $_{SB}r_{XX}$) serán más apropiadas en la aplicación de la fórmula [5.22]. Schmidt y Hunter (1996) examinaron 26 casos concretos de investigación, mostrando cual sería la corrección más apropiada en cada uno de ellos y cuales las consecuencias de no hacerla o de realizar correcciones inapropiadas.

Por otra parte, y atendiendo ahora a las relaciones entre la longitud de un test y su fiabilidad, es lógico que si la fiabilidad influye directamente en el coeficiente de validez, la longitud del test (y, en su caso, del criterio) influya también en r_{XY}, aunque de modo indirecto. Para estimar el efecto que un cambio en la longitud del test o del criterio tiene sobre el coeficiente de validez, es suficiente con estimar el coeficiente de fiabilidad del test o del criterio alargados (aplicando la fórmula de Spearman-Brown) e incorporar estos valores a la ecuación [5.21].

No obstante, la TCT proporciona expresiones que calculan directamente los cambios en validez derivados de un cambio en la longitud. Así, por ejemplo, la fórmula que permite estimar el coeficiente de validez de un test alargado *n* veces (compuesto por *n* formas paralelas) es:

$$R_{XY} = \frac{r_{XY}}{\sqrt{\frac{1-r_{XX}}{n} + r_{XX}}} \qquad [5.24]$$

Donde:

R_{XY} es el coeficiente de validez del test alargado respecto al mismo criterio.
r_{XY} es el coeficiente de validez del test original.
r_{XX} es el coeficiente de fiabilidad del test original.
n es el número de veces que se alarga el test original.

La expresión [5.24] se demuestra como sigue. Sean r_{XY}, r_{XX} y r_{YY}, respectivamente, los coeficientes de validez, de fiabilidad del test y de fiabilidad del criterio. Supongamos que alargamos con formas paralelas la longitud del test, con lo cual aumentarán su coeficiente de fiabilidad (R_{XX}) y su coeficiente de validez (R_{XY}), mientras que en el criterio (que no se modifica) el coeficiente de fiabilidad es el mismo. Según las relaciones vistas anteriormente, podemos establecer las siguientes igualdades, para el coeficiente de validez del test inicial y del alargado:

$$r_{XY} = r_{V_X V_Y} \sqrt{r_{XX} r_{YY}}$$

$$R_{XY} = r_{V_X V_Y} \sqrt{R_{XX} r_{YY}}$$

Dividiendo término a término y despejando el coeficiente de validez del test alargado, obtenemos:

$$R_{XY} = \frac{r_{XY}}{\sqrt{\frac{r_{XX}}{R_{XX}}}} = \frac{r_{XY}}{\sqrt{\frac{r_{XX}}{nr_{XX} \Big/ \left(1 + (n-1) r_{XX}\right)}}} = \frac{r_{XY}}{\sqrt{\frac{1-r_{XX}}{n} + r_{XX}}}$$

Ejemplo 5.9. Estimación del coeficiente de validez de un test alargado

Supongamos que una "Escala de actitud hacia grupos ecologistas" de 30 ítems manifiesta en un grupo normativo un coeficiente de fiabilidad de 0,51 y un coeficiente de validez de 0,42. Si se duplicase le longitud de la escala, es decir si se le añadiera una forma paralela de 30 ítems, el coeficiente de validez (respecto al mismo criterio) pasaría a valer:

$$R_{XY} = \frac{0{,}42}{\sqrt{\frac{1-0{,}51}{2}+0{,}51}} = 0{,}48$$

Si de la fórmula [5.24] despejamos n, podemos estimar el número de veces que deberemos multiplicar la longitud del test para alcanzar un coeficiente de validez R_{XY} deseado:

$$n = \frac{1-r_{XX}}{\frac{r_{XY}^2}{R_{XY}^2}-r_{XX}} \qquad [5.25]$$

En caso de que el valor de n sea negativo, significa que el valor deseado no es alcanzable incrementando la longitud del test. En el caso hipotético de un test infinitamente largo o, lo que es lo mismo, de un test con máxima precisión, en la siguiente fórmula ([5.26]), R_{XX} valdría 1, y R_{XY} se podría interpretar como el máximo coeficiente de validez obtenible como resultado de mejorar la fiabilidad del test todo lo posible.

$$R_{XY} = \frac{r_{XY}}{\sqrt{\frac{r_{XX}}{R_{XX}}}} = \frac{r_{XY}}{\sqrt{\frac{r_{XX}}{1}}} = \frac{r_{XY}}{\sqrt{r_{XX}}} \qquad [5.26]$$

Ejemplo 5.10. Estimación del número de formas paralelas a añadir para alcanzar cierto valor del coeficiente de validez.

Un determinado test de 10 ítems manifiesta en un grupo normativo un coeficiente de fiabilidad de 0,4 y un coeficiente de validez de 0,35. Nos cuestionamos cuántos ítems paralelos necesitaría el test para conseguir un coeficiente de validez de 0,5. Aplicando [5.25]:

$$n = \frac{1-0{,}4}{\frac{0{,}35^2}{0{,}5^2}-0{,}4} = 6{,}7$$

Podemos comprobar a partir de estos cálculos que el coeficiente de validez de 0,5 lo conseguiremos con un test de 67 ítems (6,7 formas paralelas de 10 ítems).

Para conseguir un coeficiente de validez de 0,9, al aplicar la fórmula obtendríamos:

$$n = \frac{1-0,4}{\frac{0,35^2}{0,9^2}-0,4} = -2,4$$

Por tanto, el coeficiente de validez de 0,9 es imposible de conseguir, por mucho que incrementemos la longitud del test inicial con formas paralelas, de ahí que hayamos obtenido un valor de *n* negativo. El máximo coeficiente de validez obtenible mejorando la fiabilidad del test (alargando su longitud) es 0,55:

$$R_{XY} = \frac{r_{XY}}{\sqrt{r_{XX}}} = \frac{0,35}{\sqrt{0,4}} = 0,55$$

El tamaño, la representatividad y la variabilidad de la muestra en el test y en el criterio

Para la estimación del coeficiente de validez es importante que la muestra donde se obtiene sea representativa de la población y de tamaño suficiente. Especialmente importante es la variabilidad que manifiesta en *X* e *Y*. De forma parecida a las relaciones que entre la varianza del grupo en el test y el coeficiente de fiabilidad (capítulo 3), el coeficiente de validez de un test respecto a un criterio es tanto más elevado cuanto mayor es la varianza de la muestra en ambos. Por ejemplo, un test de aptitud para la venta tendrá un coeficiente de validez mayor en una muestra de la población general (donde habrá heterogeneidad respecto a la aptitud por ser vendedor) que en una muestra de vendedores experimentados (seguramente obtendrían todos puntuaciones elevadas, y por tanto sería un grupo más homogéneo).

Ejemplo 5.11. Reducción del coeficiente de validez a consecuencia de la reducción en la variabilidad de la muestra

Tomamos como ejemplo los datos obtenidos por simulación de la muestra de 1.000 estudiantes, en la que se intentaba predecir el promedio de las calificaciones obtenidas en el primer curso del grado de Psicología (*Y*) a partir de las puntuaciones en un hipotético test de admisión al centro *(X)*. En la Tabla 5.7 podemos observar las consecuencias que una reducción en la variabilidad de la muestra tendrían para el coeficiente de validez. Si para calcular el coeficiente de validez dispusiésemos solamente de las puntuaciones en el criterio de los estudiantes que superaron el examen de admisión, que en nuestro ejemplo serían quienes obtuviesen 30 o más puntos en el test, el valor del coeficiente de validez sólo llegaría a 0,43. Nótese que al aplicar un punto de corte en el test, aprobar el examen, no sólo se reduce la variabilidad en el test, ya que también se reduce la variabilidad en el criterio;

si la correlación entre test y criterio es elevada también se excluirán sujetos que tendrían puntuaciones bajas en *Y*.

Tabla 5.7. Coeficientes de validez calculados para el total de la muestra y para el subgrupo de estudiantes que aprobarían el examen de admisión

Tamaño de la muestra	S_X	S_Y	r_{XY}
Total (N = 1.000)	4,886	0,973	0,532*
Estudiantes que aprobaron el examen (N= 520)	3,059	0,894	0,433*

*p < 0,05

La variable sobre la que se realiza la selección, en nuestro ejemplo el test, se denomina *directa o explícitamente selectiva* y la variable cuya variabilidad se ve reducida indirectamente, en nuestro ejemplo el criterio, se denomina *incidental o indirectamente selectiva*.

En la medida que el poder predictivo de un test respecto a un criterio depende de r_{XY}, habrá que considerar la variabilidad del grupo donde se ha obtenido. En ocasiones, por ejemplo en contextos de selección, es inevitable calcular el coeficiente de validez en una muestra de variabilidad reducida, ya que sólo de los admitidos podrá conocerse su rendimiento en el criterio *Y*. Nos encontramos entonces con un *problema de restricción del rango de variación*, puesto que nuestro interés era conocer el coeficiente de validez para el grupo completo de aspirantes que se presenta al proceso de selección. Si calculamos el coeficiente de validez de la única forma posible, esto es, correlacionando las puntuaciones de las personas seleccionadas en el test y en el criterio, el coeficiente de validez que se obtenga no nos indicará la capacidad de las puntuaciones en el test para predecir el rendimiento de los aspirantes al puesto.

Las fórmulas de Pearson-Lawley permiten corregir por restricción de rango en función de la información disponible (p. ej.: que no se conozcan las puntuaciones en el test para el grupo no seleccionado, que esto ocurra en el criterio o que se haya hecho la selección por una tercera variable). Cada escenario concreto requiere la aplicación de la fórmula adecuada. Una exposición completa puede encontrarse en Sackett y Yang (2000). Para aplicarlas hay que asumir que la recta de regresión es la misma en el grupo completo y en el reducido, así como la homocedasticidad de los errores de pronóstico en ambos grupos. Es decir:

$$b_1 = B_1 \Rightarrow r_{XY}\frac{s_Y}{s_X} = R_{XY}\frac{S_Y}{S_X}$$

$$s_{Y-Y'} = S_{Y-Y'} \Rightarrow s_Y\sqrt{1-r_{XY}^2} = S_Y\sqrt{1-R_{XY}^2}$$

Donde las letras minúsculas se refieren al grupo en el que se conocen todos los datos (normalmente el grupo de rango reducido) y las letras mayúsculas al grupo donde falta alguna información (normalmente el grupo completo). Partiendo de los supuestos anteriores, y conociendo la varianza de una de las variables en los dos grupos. se puede estimar

el coeficiente de validez desconocido. Por ejemplo, para la situación más común, con dos variables, test (X) y criterio (Y), y realizándose una selección explícita sobre el test, el coeficiente de validez puede estimarse mediante la expresión [5.27].

Efectivamente, al despejar el valor S_Y en la igualdad de los coeficientes de regresión:

$$S_Y = \frac{r_{XY} s_Y S_X}{R_{XY} s_X}$$

Y si este valor se sustituye en la igualdad de los errores típicos de estimación:

$$s_Y \sqrt{1 - r_{XY}^2} = \frac{r_{XY} s_Y S_X}{R_{XY} s_X} \sqrt{1 - R_{XY}^2}$$

Elevando al cuadrado y simplificando, la igualdad queda como:

$$\frac{1 - R_{XY}^2}{R_{XY}^2} = \frac{s_X \left(1 - r_{XY}^2\right)}{S_X^2 r_{XY}^2}$$

Finalmente, despejando R_{XY}:

$$R_{XY} = \frac{S_X r_{XY}}{\sqrt{S_X^2 r_{XY}^2 + (1 - r_{XY}^2) s_X^2}} \qquad [5.27]$$

Ejemplo 5.12. Cálculo del coeficiente de validez tras aplicar la corrección por restricción de rango, siendo el test la variable explícitamente selectiva

Un test X se ha utilizado como prueba de selección para un determinado puesto de trabajo. La varianza de las puntuaciones obtenidas en el test en el grupo completo de aspirantes fue 12 y en el grupo de admitidos fue 6. En este último grupo su correlación con el criterio fue 0,72 y la varianza de las puntuaciones en el criterio 7 ¿Cuál estimamos que sería el coeficiente de validez del test en el grupo completo de solicitantes?

Sustituyendo en la expresión [5.27]:

$$R_{XY} = \frac{\sqrt{12} \cdot 0{,}68}{\sqrt{12 \cdot 0{,}68^2 + (1 - 0{,}68^2)6}} = 0{,}79$$

Que es superior al que se obtuvo en el grupo de admitidos (0,72).

En estas aplicaciones hay que ser cauto, ya que el supuesto de homocedasticidad de los errores de pronóstico suele ser falso, es decir, la varianza de dichos errores suele diferir para grupos con diferente nivel de rasgo. Si en el grupo seleccionado la varianza de los errores de pronóstico fuese menor, el coeficiente de validez corregido estará sobrestimado.

En la página Web de Paul Barret (http://www.pbarret.net) se puede obtener un programa específico para calcular con comodidad, en las distintas situaciones posibles, las correcciones de los coeficientes de validez por restricción de rango. Este programa también ofrece la posibilidad de calcular el coeficiente de validez corregido por atenuación.[9]

Evidencias basadas en los procesos de respuesta a los ítems

Un modo de obtener información sobre las inferencias que podemos realizar con las puntuaciones de un test es analizar los procesos de respuesta que los sujetos deben realizar para obtener dichas puntuaciones. Para ello, se requiere de un modelo explicativo (una teoría psicológica sustantiva) de dichos procesos de respuesta, que debería guiar el proceso de construcción del test, y que debería servir para predecir el diferente rendimiento en los ítems. Borsboom, Mellenbergh y van Heerden (2004) defienden que el análisis de las evidencias sobre la validez de las puntuaciones obtenidas en un test es un tema que atañe más al proceso de construcción del instrumento que a los estudios de covariación realizados a posteriori, tan tradicionalmente enfatizados para obtener evidencias sobre la estructura interna o sobre las relaciones con otras variables. Para ellos, al construir un test, debe tenerse una idea clara de cómo diferentes niveles en el atributo que se pretende medir deberían llevar a distintas puntuaciones empíricas; esto sólo puede hacerse partiendo de una teoría psicológica muy sólida sobre los procesos de respuesta a los ítems. Embretson y Gorin (2001) muestran un buen ejemplo de cómo se puede utilizar el análisis de los procesos de respuesta para obtener evidencias sobre la validez de las puntuaciones de un test diseñado para medir capacidad espacial. El análisis de los procesos permitió distinguir dos tipos de ítems: aquellos que para su resolución requerían rotación mental y los que podían resolverse sin necesidad de ésta, simplemente por un procesamiento perceptual general. Los segundos serían menos válidos para evaluar la capacidad espacial de las personas.

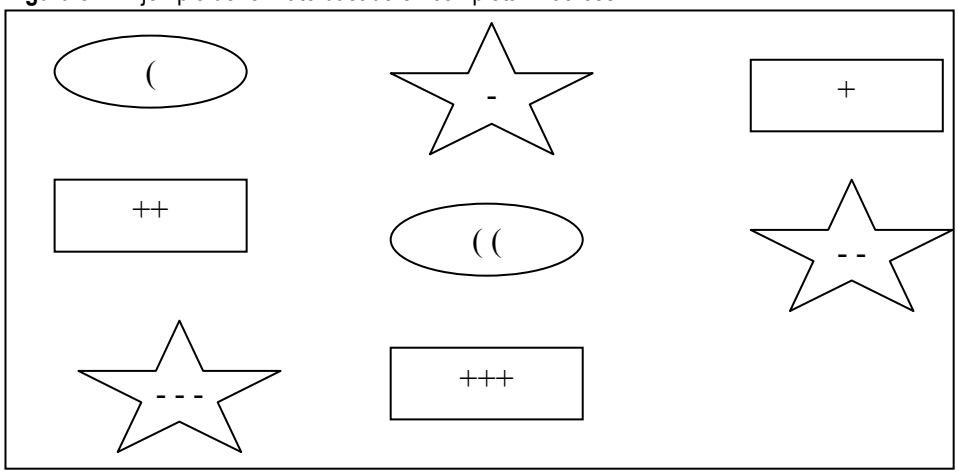
Algunos autores (por ejemplo, Bejar, 2002) emplean la denominación de *tests basados en modelos* para referirse el diseño de instrumentos de evaluación guiados por una teoría psicológica sobre el procesamiento de respuestas. Embretson (2002) propone la expresión *representación del constructo* para referirse al conjunto de procesos, estrategias y estructuras de conocimiento que están implicados en la resolución de los ítems; esta autora señala que la investigación previa de los psicólogos cognitivos es muy relevante para conocer qué variaciones en los estímulos deben establecerse para conseguir que los ítems tengan diferente nivel de demanda cognitiva, y por tanto diferente dificultad. Para esta autora, el diseño de tests desde un enfoque cognitivo debería seguir el siguiente procedimiento, ejemplificado con el trabajo realizado para elaborar un test de razonamiento abstracto:

---

[9] Johnson y Ree (1994) desarrollaron el programa RANGEJ que permite calcular la restricción de rango para el caso de múltiples variable predictoras.

1. *Especificar los objetivos de la medición*. Por ejemplo, la medición del razonamiento abstracto como componente esencial de la inteligencia fluida. Se trata de identificar el tipo de tareas y las características que deben manipularse para alterar la exigencia cognitiva que se plantean. Concretando, en el test de razonamiento abstracto deben establecerse ítems relativamente independientes de los conocimientos previos de las personas. Atendiendo a las experiencias con tests previos (por ejemplo, el Test de Raven) y a la investigación realizada sobre el procesamiento de este tipo de tareas, se eligió un formato de "completar matrices" como el ofrecido en la Figura 5.2.

Figura 5.2. Ejemplo de formato basado en completar matrices



2. *Establecer un modelo de procesamiento*, donde se indiquen tres cosas: en primer lugar, los procesos, estrategias y estructuras de conocimiento implicados; en segundo lugar, deben operacionalizarse (cuantificarse) las características de los ítems que influyen en su procesamiento; en tercer lugar, deben establecerse previsiones sobre la influencia de la manipulación de las características sobre las propiedades psicométricas de los ítems, por ejemplo sobre su dificultad. En el caso concreto del test de razonamiento abstracto, se siguió el Modelo de Procesamiento de Matrices de Carpenter, Just y Shell (1990), que básicamente establece un procesamiento serial como el siguiente: codificar las dos primeras figuras de la primera fila, determinar los elementos correspondientes, comparar los atributos de los elementos, inferir una regla inicial de relaciones, codificar la tercera figura, comparar sus elementos con los de las figuras iniciales, inferir si la regla inicial es correcta o debe proponerse otra, repetir el proceso con el resto de las filas y con las columnas. Respecto a las características de los ítems que influyen en su procesamiento, se establecieron diferentes niveles de dificultad previsible atendiendo a los contenidos de las figuras y a las reglas que gobiernan las relaciones entre ellas; por ejemplo, es más fácil resolver un ítem donde los símbolos internos son siempre los mismos (o simplemente no aparecen) que otro ítem que incluye símbolos diversos y de carácter más abstracto; será más complicado un ítem donde las figuras son muy parecidas (en el caso de que unas sean distorsiones ligeras de otras) que otro ítem con figuras

claramente diferenciadas; también influyen las reglas que gobiernan las relaciones (por ejemplo, no sería fácil descubrir que el tercer elemento de una fila o columna se obtiene restando los anteriores); además, será más complicado resolver un ítem donde se establecen varias reglas de relaciones entre las figuras que en otro gobernado por una regla simple que exige menor carga memorística. Se realizaron varios estudios empíricos con el Test de Raven para comprobar cuáles de estas características incidían en la dificultad de los ítems. Lo importante en este punto del proceso es que se dispone ya de un modelo de procesamiento que concreta las variables que deben manipularse para generar ítems con diferente demanda cognitiva.

3. *Generar ítems*, de tal forma que las variaciones en su estructura representen variaciones en los procesos de respuesta. A partir de los resultados de los estudios realizados con el Test de Raven, comienza propiamente el diseño del nuevo test. Se establecieron las características fisicas de las figuras que debían manipularse y el número de reglas aplicadas en las relaciones entre figuras. Con un programa informático se generaron todos los ítems posibles (150 en total) que combinaban las características establecidas y el número de reglas.
4. *Evaluar empíricamente las previsiones del modelo* sobre el rendimiento de los sujetos en los ítems, así como establecer los oportunos estudios de validez. Varios estudios empíricos mostraron la influencia que tenían las variables consideradas en la fase de elaboración de los ítems sobre su dificultad empírica. Por ejemplo, la cantidad de reglas incluidas, el grado de abstracción de las figuras y otras características perceptivas de los ítems explicaron un 79 % de la varianza de los parámetros de dificultad de los ítems, estimados mediante el Modelo de Rasch, y un 77% de las latencias de respuesta o tiempo tardado en resolver los ítems. En cuanto a otras evidencias de validez, se comprobó que todos los ítems saturaban en un único factor y que también los ítems del Test de Raven saturaban en dicho factor.

Otro ejemplo de evidencias sobre los procesos de respuesta lo describe Hornke (2002) en un test de rotación de figuras, donde se manipula la cantidad de elementos a procesar, si las figuras son bi o tridimensionales, el ángulo de la rotación y el número y tipo de rotaciones (de derecha a izquierda, de arriba a abajo,..). En el mismo capítulo, este autor describe un test de memoria visual en el que los ítems son planos de una ciudad donde aparecen determinados iconos para representar ciertos servicios públicos, manipulándose en cada caso la cantidad de iconos, su tamaño o el nivel dispersión en el mapa.

Vemos entonces que en este tipo de enfoque no sólo se miden las respuestas del sujeto a los ítems, sino que se consideran los pasos intermedios ejecutados para obtener dichas respuestas. Por otra parte, el conocimiento sobre los componentes requeridos para la respuesta correcta de los ítems no sólo es importante para la obtención de evidencias de validez; este modo de proceder permite una información diagnóstica mucho más completa, pues es posible conocer los componentes en los que los examinados tienen dominio y aquellos en los que presentan dificultades.

Se han desarrollado modelos de TRI específicos para analizar la incidencia de los diversos procesos establecidos desde el marco teórico. Así, el Modelo Logístico Lineal de Rasgo Latente (LLTM, Fischer, 1973) fue el primer modelo componencial desarrollado y el que ha sido empleado con mayor frecuencia. En los modelos componenciales se entiende que para ejecutar cierta tarea es necesario desarrollar una serie de componentes o procesos (ya sea secuencial o concurrentemente). El modelo LLTM permite estimar, además

de los niveles de rasgo de las personas y la dificultad de los ítems, la contribución de los diferentes componentes a dicha dificultad.

Ejemplo 5.13. Aplicación del modelo LLTM de Fisher a un test de aritmética

Romero, Ponsoda y Ximénez (2008) analizaron un test de aritmética mediante el modelo LLTM. Este test ha sido diseñado para niños que acaban de aprender el concepto de suma y resta con números enteros. Contiene 32 ítems de opción múltiple con 4 alternativas de respuesta, y se pide la adición o sustracción entre dígitos enteros. Un ejemplo de ítem es: **(−6) + (3)** = *a)* 9 *b)*3 *c)*−3 *d)*−9. Los autores proponen 6 operaciones o componentes: O1: Adición entre números naturales (a+b); O2: Sustracción entre números naturales (a−b) cuando a>b; O3: Identificación del componente mayor en valor absoluto y planteamiento de resta del menor al mayor; O4: Cambiar las posiciones de a y b; O5: Determinar el signo (positivo o negativo) del resultado; O6: Convertir la sustracción en adición y cambiar el signo al segundo dígito. A modo de ejemplo, el ítem (−6)+(5) debería requerir aplicar primero O3: 6−5, luego O2: 1 y finalmente O5: −1.

Al estimar los parámetros del modelo, se obtuvo que 4 de los 6 pesos (componentes) resultaron significativos, por lo tanto estas operaciones contribuyen a la dificultad de los ítems. Por ejemplo, se encontró que la operación que contribuía en mayor medida a la dificultad de los ítems era O6; esto era de esperar, pues se refiere a un proceso doble que implica no sólo cambiar el operador resta por suma sino también cambiar el signo del segundo dígito; por ejemplo (−a)−(−b)=−a+b.

Modelos como el LLTM representan , además, la base psicométrica de la *generación automática de ítems* (GAI). Si conocemos las variables que intervienen en el procesamiento de los ítems, puede establecerse un método para construir todo el universo posible de ítems gobernado por dichas variables. La GAI consiste en la construcción de bancos de ítems mediante algoritmos; se establece un conjunto de reglas explícitas, susceptibles de programarse en un ordenador, que determinan cómo deben construirse los ítems y predecir la dificultad de cada uno a partir de los componentes involucrados; sería posible, por lo tanto, la aplicación de ítems sin previa calibración (ver p. ej., Revuelta y Ponsoda, 1998b).

En las últimas décadas se intenta estrechar la distancia entre los modelos cognitivos y los modelos psicométricos. Información más específica sobre los diferentes tests, los modelos en que se sustentan y los estudios realizados para obtener evidencias de validez puede consultarse en Irvine y Kyllonen (2002).

Evidencias basadas en las consecuencias de la aplicación del test

Resulta cada vez más usual la aplicación de tests psicológicos y educativos en determinados marcos institucionales y organizacionales. Por ejemplo, se aplican tests de conocimientos o competencias escolares para evaluar el nivel alcanzado por los estudiantes en un determinado ciclo de enseñanza. Se emplean tests de diverso tipo en procesos de selección de personal con objeto de predecir el rendimiento laboral de los aspirantes. En contextos de evaluación de programas, los tests sirven como instrumentos de medida de los cambios producidos por la intervención social efectuada. En todos estos escenarios, la mera aplicación de tests puede tener consecuencias sociales diferentes al propósito fundamental que se pretende con la aplicación, lo que ha llevado a incorporar en la última edición de los *Standards* (AERA, APA, NCME, 1999) la necesidad de aportar evidencias sobre la denominada *validez consecuencial*, es decir, el análisis de las consecuencias intencionadas y no intencionadas que se derivan de la aplicación de tests en determinados contextos de evaluación. La revista *Educational Measurement: Issues & Practice* publicó dos números monográficos sobre el tema en 1997 y 1998. Gran parte de la sensibilidad actual a las consecuencias del uso de los tests tiene que ver con la legislación estadounidense *No Child Left Behind*, que ha llevado a la aplicación masiva de tests para la evaluación de conocimientos y destrezas de los escolares dentro de una política para favorecer la "rendición de cuentas" de los centros educativos y mejorar la enseñanza y el aprendizaje de los estudiantes. La utilización de tests con altas consecuencias para los evaluados (*high stakes testing*) que se emplean, por ejemplo, para acreditaciones profesionales en USA, también ha incidido en el interés por este problema.

Las consecuencias que puede tener la aplicación de tests de conocimientos o destrezas en contextos de evaluación institucional, tal como se realiza por ejemplo en diversas comunidades autónomas españolas en niveles de Educación Primaria y Secundaria, son muy diversas. Puede llevar a que determinados centros educativos adiestren específicamente a los estudiantes en los contenidos que se van a evaluar, produciéndose un "estrechamiento curricular" con objeto de que sus estudiantes rindan mejor en los tests y el colegio salga "mejor parado" en comparación con los centros del entorno (una consecuencia negativa denominada en inglés *test pollution*) o puede servir para que los claustros de profesores analicen el modo de mejorar el proceso instruccional en las asignaturas donde sus estudiantes no manifiestan un buen rendimiento (una consecuencia positiva). Como los resultados de la evaluación son públicos, pueden influir en la elección del centro por las familias para la educación de sus hijos. Algunos centros con elevada tasa de niños inmigrantes pueden aparentemente rendir peor que otros si no se asegura que los tests no manifiestan funcionamiento diferencial contra este tipo de minorías. El nivel previo de los estudiantes, determinado en parte por variables familiares y sociales, tampoco será independiente del rendimiento obtenido, con lo que los resultados no pueden atribuirse exclusivamente a la acción educativa. En algunos países, como Estados Unidos, parte de la subvención pública de los colegios depende del rendimiento conseguido por los estudiantes en tests de conocimientos escolares, estableciéndose sanciones a los centros cuyos estudiantes no alcancen determinadas competencias académicas. Además, incluso se proponen modificaciones en la política educativa, en el diseño curricular o en la retribución de los profesores, a partir de los resultados de las evaluaciones. Algunos estudios realizados en Estados Unidos re-

velan que muchos profesionales de la educación han perdido motivación laboral, que se sienten realmente presionados para alcanzar los estándares y que no perciben mejoras relevantes en el proceso de enseñanza-aprendizaje. Además, la falta de motivación de los estudiantes al responder a los tests (hartos de que todos los años se les pida algo sobre lo que no perciben consecuencias académicas) representa un importante problema que afecta a la validez de las puntuaciones obtenidas bajo este tipo de condiciones.

En dos recientes trabajos (Padilla, Gómez, Hidalgo y Muñiz, 2006, 2007) se profundiza sobre este tema, revisando las diferentes posturas que mantienen los psicómetras, analizando las dificultades que conlleva el estudio de las consecuencias del uso de los tests, y delimitando el tipo de consecuencias de las que debe informarse en el proceso de validación de las puntuaciones. Ha habido una fuerte polémica con autores a favor de la consideración de estas evidencias (p. ej.: Cronbach, 1988; Messick, 1980) y en contra (p. ej.: Boorsboom, et al. 2004; Popham, 1997). Para los primeros es fundamental saber si el test puede tener consecuencias sociales en contextos donde ciertos grupos resulten sistemáticamente desfavorecidos; para ellos, hay que recoger información no solo sobre la interpretación de las puntuaciones, sino también sobre el uso justificado de las mismas. Los segundos consideran que no se está hablando de evidencias empíricas sobre las inferencias que pueden realizarse con las puntuaciones y, por tanto, creen que no debería incluirse este tipo de evidencias en el proceso de validación. En los *Standards* se plantea la necesidad de analizar explícitamente las consecuencias del uso de los tests, diferenciando entre aquéllas que tienen que ver con su validez y las que, aun siendo importantes, caen fuera de este ámbito. Si la evidencia empírica permite mantener las interpretaciones, la decisión final sobre el uso del test puede tener en cuenta otras consideraciones sociales o políticas que ya no formarían parte del proceso de validación.

Dada la dificultad que entraña la comprobación de todo tipo de consecuencias sociales que pueden seguirse de determinadas aplicaciones, algunos autores recomiendan centrarse en las que pueden derivarse de una limitada representación del constructo o de la presencia de factores irrelevantes al constructo. En un reciente artículo, Nichols y Williams (2009) describen ambos tipos de consecuencias con dos casos concretos. En relación a la infra-representación del constructo, describen la preocupación que tienen algunas universidades norteamericanas porque, a raíz de aplicar un test para la admisión muy cargado en conocimientos científicos básicos, los candidatos se preparan muy específicamente en cursos sobre Ciencia y no en otro tipo de conocimientos y destrezas relevantes para ese tipo de estudios. En cuanto a la presencia de factores irrelevantes al constructo, se refieren a los sesgos de corrección de ensayos debidos al diferente grado de dureza establecido por los correctores cuando se escriben a mano o cuando se escriben con el ordenador; parece que en éstos últimos los correctores son más estrictos.

¿Qué procedimientos o técnicas podemos aplicar para aportar evidencias sobre las consecuencias de las aplicaciones de los tests? Resulta evidente que es muy difícil anticipar todo tipo de consecuencias y aportar evidencias empíricas sobre las mismas. Sin embargo, algunas orientaciones, tomadas principalmente de la experiencia en la aplicación de tests de conocimientos y competencias académicas en contextos escolares, pueden ser las siguientes:

1. En la construcción de un test podemos justificar y analizar la representación del constructo. Por ejemplo, los tests no pueden incluir contenidos muy limitados que impidan generalizar el rendimiento a los objetivos de aprendizaje planteados para el nivel edu-

cativo y que permitan un mejor rendimiento a través de un entrenamiento específico a los tests.

2. Puede ser útil comprobar si la estructura interna del test, aplicado en un contexto determinado, se mantiene en una nueva aplicación del mismo en otras condiciones. Por ejemplo, existe evidencia de que la estructura interna del Modelo de Cinco Factores de la Personalidad de ciertos tests no se mantiene cuando se aplican en procesos de selección de personal, donde los aspirantes han sido orientados a proporcionar una buena imagen en sus respuestas.
3. Por otra parte, disponemos de procedimientos y técnicas para estudiar el *sesgo* y el *impacto adverso*, temas que se abordarán en este libro en el capítulo 13 y que representan algunas de las consecuencias indeseables relacionadas con la presencia de factores irrelevantes al constructo.
4. Determinadas consecuencias pueden evaluarse mediante la aplicación de cuestionarios o entrevistas a las personas que pueden verse afectas por la aplicación de los test. Por ejemplo, en algunos países se pregunta a los profesores, directores, estudiantes y familias sobre sus opiniones respecto a la utilidad y consecuencias de los procesos de evaluación educativa. Un excelente trabajo sobre las opiniones de los profesores respecto al impacto de la evaluación educativa que se realiza en Estados Unidos puede consultarse en la siguiente dirección: http://www.education.uiowa.edu/cea/__documents/Consequential_Validity_NCME_2006.pdf
5. Pueden realizarse también investigaciones empíricas para estudiar determinados efectos. Por ejemplo, diseños longitudinales donde se analicen los cambios producidos por los programas de evaluación educativa en el rendimiento de los estudiantes, en las prácticas educativas o en otro tipo de variables dependientes. También pueden estudiarse longitudinalmente los efectos del entrenamiento específico sobre tests similares a los que se aplican.

Nichols y Williams (2009) delimitan las responsabilidades de los profesionales que hacen los tests de los responsables de las aplicaciones. En general, los primeros deberían anticipar consecuencias inmediatas o persistentes, pero no son los responsables de aplicaciones inadecuadas o de los efectos a largo plazo.

Evolución histórica del concepto de validez

Acabamos de desarrollar la concepción actual de validez y de mostrar distintos procedimientos utilizados para obtener evidencias sobre la validez de las puntuaciones pero, como señalábamos al principio del capítulo, el concepto de validez ha cambiado mucho a través del tiempo ¿Cómo hemos llegado al concepto actual de validez? ¿Qué cambios se han producido en su definición? ¿Por qué han tenido lugar? Intentaremos responder a estas cuestiones en los siguientes párrafos. Kane (2006a) proporciona una detallada exposición de esta evolución.

Una primera época en la conceptualización de la validez se extiende desde 1920 hasta 1950 y podría resumirse como un *modelo de validez referida a un criterio*. Este período está dominado por una mentalidad práctica y operacionalista. Los tests servían para medir aquella variable *observable* con la que presentaban una alta relación. Lo importante era

que el test tuviese la capacidad de predecir un criterio externo (Gulliksen, 1950). Este modelo es simple y eficaz si podemos disponer de un criterio plausible. Esto ocurre, por ejemplo, en muchos contextos aplicados donde el objetivo es predecir el rendimiento en un curso o un trabajo. Las medidas de la ejecución real en esas tareas se pueden usar como criterio. De hecho, ésta es todavía la aproximación a la validez preferida en este tipo de aplicaciones.

Durante esta primera etapa también se buscaron argumentos sobre la validez de los tests mediante la revisión de sus contenidos por jueces expertos, con objeto de decidir si los elementos del test eran relevantes y representativos. El análisis del contenido era, y como hemos visto sigue siendo, frecuentemente aplicado en las medidas de rendimiento académico. Su subjetividad es su principal limitación, ya que la evaluación recae sobre la opinión de unos jueces. Además, algunos autores (como Messick,1989) consideran que juega un limitado papel en la validación, ya que no proporciona evidencia directa sobre las inferencias que se pueden hacer a partir de las puntuaciones en el test.

Por lo tanto, a principios de 1950 el estudio de la validez estaba basado en la capacidad para predecir un criterio y en el análisis del contenido del test. Pero ¿qué hacer en situaciones donde no es posible disponer de un buen criterio? ¿Cuál sería el criterio para medir la Inteligencia o la Creatividad? En los años 50 ocurrió un cambio importante. La APA publicó en 1954 sus primeras normas sobre los tests ("*Technical Recommendations for Psychological Tests and Diagnostic Techniques*"), en las que se reconoce que la validación en base a un criterio no siempre es posible; en estas normas se plantea la necesidad de obtener evidencias para justificar las interpretaciones que hacían los psicólogos clínicos. Surge así, en el período comprendido entre 1955 y 1989, un nuevo *modelo de validez basado en el concepto de constructo*. Los constructos se definían como atributos no observables que se reflejaban en las respuestas a un test. Una contribución esencial en esta etapa es el artículo de Cronbach y Meehl (1955), que ha sido probablemente el trabajo que más ha influido en nuestra concepción actual de la validez. Los autores afirmaban que aunque en un test se hubiese llevado a cabo una validación de contenido o referida a un criterio, era deseable, para la mayoría de los casos, la determinación del constructo medido. La validación de constructo suponía apoyarse en una red nomológica, es decir, en un sistema que representase las relaciones existentes entre los constructos objeto de estudio a partir de sus manifestaciones observables, y que permitiese formular hipótesis empíricamente contrastables. Desde esta nueva conceptualización se considera que la validación es un proceso mucho más complejo cuya efectividad depende de la disponibilidad de un modelo teórico previo, de una teoría bien definida. Por su parte, Campbell y Fiske (1959), ofrecieron un procedimiento empírico para la validación de constructo basado en el análisis de las matrices MRMM.

La nueva conceptualización se recogió muy lentamente en las sucesivas ediciones de los *Standards* (versiones de 1966 y 1974). En ellos se consideraba que la validación de constructo era una de las posibles aproximaciones al estudio de la validez, cuando no existía un criterio aceptable. Además, se distinguían tres tipos de validez: validez referida a un criterio (englobaba la validez concurrente y predictiva), validez de contenido y validez de constructo. Se instauró así la denominada concepción trinitaria de la validez, todavía hoy presente en la mente de algunos profesionales.

A finales de los años 70 había dos tendencias opuestas en el desarrollo de la teoría de la validez. Por un lado, el interés en aclarar la clase de evidencias necesarias para validar

particulares interpretaciones y usos de las puntuaciones en los tests. Por otro lado, la necesidad percibida de desarrollar un concepto unitario de validación.

Los *Standard*s de 1985 intentaron resolver esta tensión reconociendo la validez como un concepto unificado y reconociendo también que diferentes tipos de evidencia eran necesarias para diferentes tipos de interpretaciones. Se mantuvo la distinción entre validez de criterio, de contenido y de constructo. Eso sí, ya no se consideraban distintos tipos de validez, sino distintos tipos de evidencias que eran necesarias para diferentes tipos de interpretaciones. Sin embargo, los teóricos de la validez (p. ej., Cronbach o Messick) defendían una aproximación mas unificada y expresaban su inquietud por la tendencia a emplear diferentes métodos de validación para diferentes usos de las puntuaciones: el modelo del criterio para validar decisiones de selección, el del contenido para validar tests de logro y el de constructo para proporcionar explicaciones teóricas.

En la segunda mitad de los años 80 se adoptó una concepción amplia de la validez de constructo, tratando de establecer un marco de trabajo unificado, que englobaba también las evidencias sobre el contenido y sobre el criterio. Desde esta perspectiva se insistía en la necesidad de disponer de teorías que propusiesen interpretaciones de las puntuaciones, así como justificarlas después de desarrollar auténticos programas de investigación (y no un único estudio empírico). Sin embargo, la nueva concepción no establecía guías; se convirtió en un "cajón de sastre" donde cabía casi cualquier tipo de evidencia. Ello puede explicarse, por un lado, por la carencia de teorías "fuertes" en Psicología. En ausencia de estas teorías, la validez de constructo tiende a ser muy abierta. Si todos los datos son relevantes para la validez ¿por dónde empezar?, ¿cuánta evidencia es necesario acumular? Por otro lado, esta confusión vino alentada porque, en definitiva, los tres tipos de evidencia coincidían con la estructura trinitaria de los tipos de validez.

En la edición más reciente de las normas sobre los tests (la edición de1999), que recoge el concepto de validez actualmente dominante, se establecen algunas aclaraciones importantes:

1. Se enfatiza el carácter unitario de la validez y se rechazan las tres categorías tradicionales de validez. La validación es una evaluación unificada de la interpretación, no simplemente un conjunto de técnicas.
2. Se destaca la centralidad de la validez de constructo en el proceso de validación. Pero se adopta una óptica más general, para entender el constructo no ya exclusivamente como un atributo teórico sino como cualquier característica medida por un test. Se pretende una definición clara y detallada de las interpretaciones propuestas y también la consideración de las interpretaciones alternativas.
3. Se añaden dos nuevos tipos de evidencias: las basadas en el proceso de respuesta a los ítems de un test y en las consecuencias sociales del proceso de aplicación del test.

Aunque esta es la concepción dominante en la actualidad, recientemente, Borsboom y sus colaboradores (Borsboom, Mellenberg y van Heerden, 2004; Borsboom, 2006) la han sometido a fuertes críticas. Consideran que la teoría actual sobre la validez ha fallado, ya que nos ha dejado con la impresión de que cualquier asunto relacionado con los tests es relevante para el problema de la validez; esto impide plantear estudios de validación realmente eficaces. Según estos autores, la validez no es un concepto complejo, ni dependiente de redes de trabajo nomológicas, ni de consecuencias sociales. Para ellos, un test

sería válido para medir un atributo si y sólo si: (1) el atributo existe y (2) variaciones en el atributo producen causalmente variaciones en los resultados de la medición.

Esto significaría, por ejemplo, que las correlaciones entre las puntuaciones en el test y otras medidas no suponen más que una evidencia circunstancial de validez. Para Borsboom, el problema de la validez no puede ser resuelto desde técnicas o modelos psicométricos que se aplican después de obtener las respuestas. Muy al contrario, el proceso de validación tiene que estar dirigido desde una teoría sustantiva y reflejarse desde el mismo diseño del test, y no después. Este marco teórico debería explicar lo que ocurre entre los niveles de atributo y las respuestas a los ítems, algo que resulta difícil porque las redes nomológicas de las teorías psicológicas normalmente resultan ambiguas.

Esta concepción parece rompedora (al menos está suscitando mucho el debate), pero todavía es muy reciente para valorar su posible incidencia en los Standards o en la práctica real de los estudios de validación. De hecho, algunos autores, como Kane (2006b) o Sijtsma (2006), afirman sentirse cómodos en el marco teórico actual y consideran que guiarse desde una teoría sustantiva es excelente, pero que es lo que se viene haciendo desde hace mucho tiempo. Para estos autores, dado que las teorías sustantivas formales no existen, el proceso de construcción debe estar guiado por concepciones generales del atributo de interés. Un modelo teórico puede ser causal, y en este sentido predecir diferentes puntuaciones para diferentes niveles de atributo, sin que necesariamente sea formal.