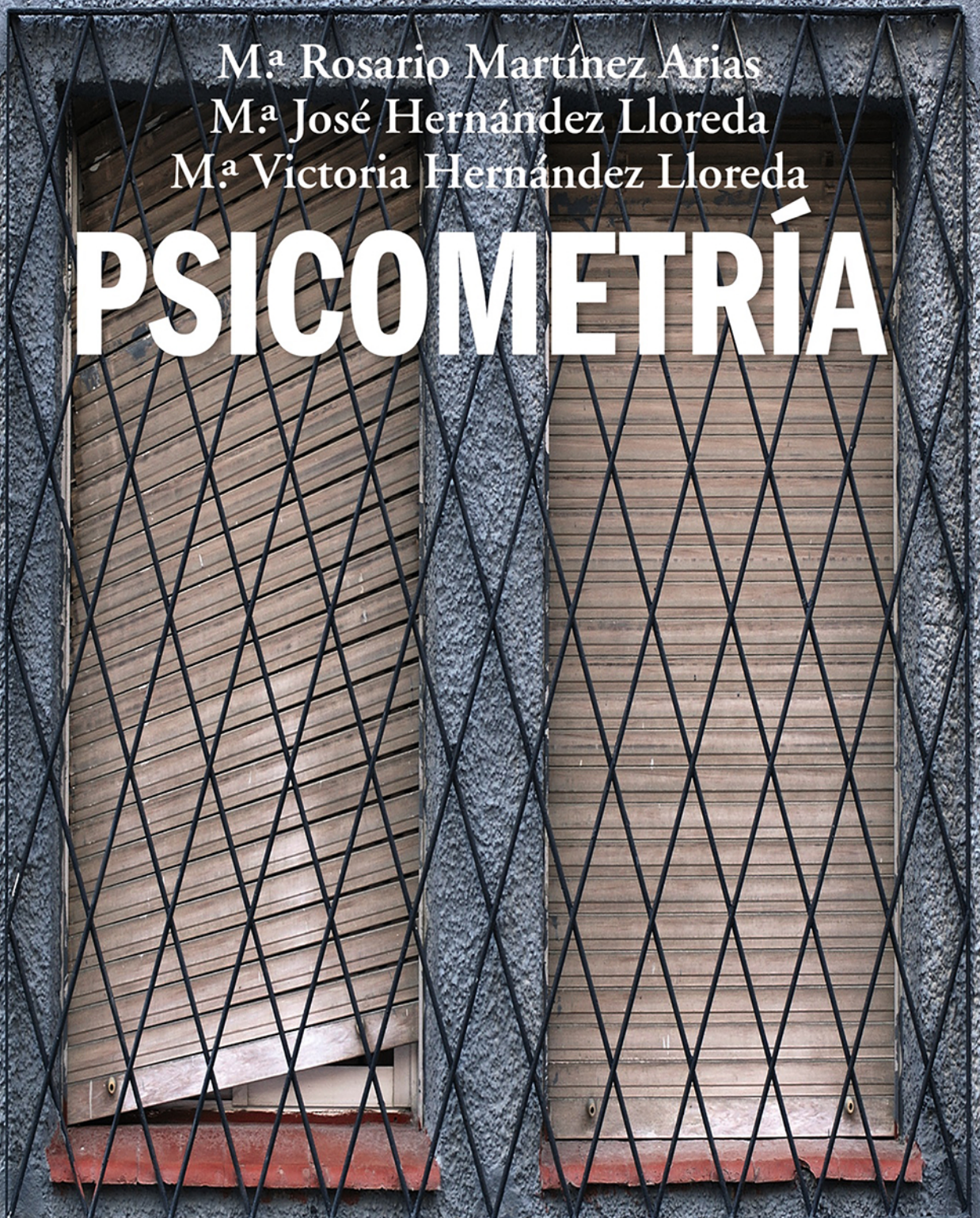
M.ª Rosario Martínez Arias
M.ª José Hernández Lloreda
M.ª Victoria Hernández Lloreda

PSICOMETRÍA



Alianza Editorial

María Rosario Martínez Arias
María Victoria Hernández Lloreda
María José Hernández Lloreda

Psicometría

Alianza Editorial

Reservados todos los derechos. El contenido de esta obra está protegido por la Ley, que establece penas de prisión y/o multas, además de las correspondientes indemnizaciones por daños y perjuicios, para quienes reprodujeren, plagiaren, distribuyeren o comunicaren públicamente, en todo o en parte, una obra literaria, artística o científica, o su transformación, interpretación o ejecución artística fijada en cualquier tipo de soporte o comunicada a través de cualquier medio, sin la preceptiva autorización.

Edición electrónica, 2014
www.alianzaeditorial.es

© María Rosario Martínez Arias, María Victoria Hernández Lloreda y María José Hernández Lloreda, 2006
© Alianza Editorial, S. A. Madrid, 2014
Juan Ignacio Luca de Tena, 15. 28027 Madrid
ISBN: 978-84-206-8859-6
Edición en versión digital 2014

Índice

Prólogo 13

1. Tests psicológicos y educativos: conceptos básicos, clasificación y construcción 15

1. Introducción 15

1.1. La medida en psicología y educación 15
1.2. ¿Por qué los profesionales necesitan conocer la teoría de los tests? 16

2. Concepto de test. Algunas clasificaciones 18
3. El proceso de inferencia psicométrica 20
4. Los principios del modelo psicométrico 23

4.1. Fiabilidad 23
4.2. Validez 24
4.3. Comparabilidad 24
4.4. Equidad 24

5. Los modelos de tests 25
6. La construcción de un test 26

6.1. La finalidad del test 27
6.2. Restricciones en la aplicación 28
6.3. Marco del test: características del constructo o dominio 28
6.4. Especificaciones derivadas del marco 30
6.5. Especificación del formato de los ítems. Reglas de puntuación 31
6.6. Revisión de los ítems 33
6.7. Estudio piloto 33

6.8. Estudio de campo 34
6.9. Manual del test y sus revisiones 35
7. Criterios para la valoración de un test 35

2. Teoría Clásica de los Tests (TCT) I: el modelo y la fiabilidad de las puntuaciones 37

1. El modelo de la TCT 37
2. Supuestos y deducciones del modelo 38
3. Concepto de fiabilidad 41
4. Relaciones del modelo basadas en el supuesto de paralelismo 44
5. Factores que afectan al coeficiente de fiabilidad 46
5.1. Fiabilidad y variabilidad de la muestra 46
5.2. Fiabilidad y longitud: fórmula de la profecía de Spearman-Brown 47
6. Estimación empírica del coeficiente de fiabilidad 50
7. La fiabilidad como consistencia interna 52
7.1. Procedimientos basados en la división del test en dos mitades 52
7.2. Métodos basados en las covarianzas entres los ítems 54
8. Coeficiente de fiabilidad como acuerdo entre jueces 60
9. Tests de velocidad y tests de potencia 62
10. Fiabilidad en situaciones más complejas: tests compuestos y puntuaciones diferencia 63
10.1. La fiabilidad de un test compuesto 63
10.2. Fiabilidad de las ganancias o puntuaciones diferencia 65
11. Algunas consideraciones sobre las estimaciones del coeficiente de fiabilidad 66
12. Consecuencias de los errores de medida sobre los análisis de datos 67

3. La Teoría Clásica de los Tests II: puntuaciones, análisis de elementos, consideraciones finales 69

1. Las puntuaciones en la TCT 69
1.1. Puntuaciones de los tests como suma de las respuestas a los ítems 69
1.2. Tests formados por ítems dicotómicos de elección múltiple (EM) 70
1.3. Problemas especiales con los tests de velocidad 72
2. Puntuaciones con ponderación de los ítems 73
3. La estimación de las puntuaciones en la TCT 74
3.1. Estimadores puntuales 74
3.2. Estimaciones de intervalos o bandas de puntuaciones 75
3.3. Interpretación de algunas diferencias 77
4. El análisis de los ítems en la TCT 79
4.1. El análisis de las opciones del ítem 80
4.2. Estadísticos de los ítems 82
4.3. Cálculo de los estadísticos de los ítems con programas de ordenador 97
5. Valoración de las TCT 99

4. Introducción a la Teoría de la Generalizabilidad (TG) 101

1. Introducción 101
2. Innovaciones de la TG 102
3. Los conceptos fundamentales de la TG 103

3.1. El objeto de la medición 103
3.2. Facetas y niveles 103
3.3. Puntuación observada *versus* puntuación del universo 104
3.4. Diseño experimental 104
3.5. Estudios G y D 105

4. Estudio de generalizabilidad (estudio G) 106

4.1. Modelo estadístico de la TG 106

5. Estudios de decisión 113

5.1. El coeficiente de generalizabilidad 113
5.2. Estimación de los componentes de la varianza en los estudios D 114
5.3. Decisiones absolutas y relativas 116
5.4. Los errores en la TG 117
5.5. Cálculo de los coeficientes de generalizabilidad 118
5.6. ¿Cómo se pueden obtener en la práctica los estimadores de los coeficientes de la Teoría de la Generalizabilidad? 121

6. Valoración general de la TG 122

5. Teoría de la Respuesta al Ítem (TRI) I: conceptos básicos y modelos para ítems dicotómicos 123

1. Introducción 123
2. Principales diferencias entre la Teoría Clásica de los Tests y la Teoría de la Respuesta al Ítem 126
3. Características generales de los modelos de la TRI 127

3.1. Aspectos comunes a los distintos modelos 127
3.2. Principales ventajas de los modelos de la TRI 130

4. Los supuestos de la TRI 131

4.1. Unidimensionalidad (dimensionalidad en general) 131
4.2. Independencia local 132

5. Los parámetros de la Curva Característica del Ítem (CCI) o Función de Respuesta al Ítem (FRI) 134
6. La forma de los modelos de la CCI 140

6.1. Modelos logísticos 141

7. La elección de un modelo para ítems dicotómicos 153

6. Teoría de la Respuesta al Ítem (TRI) II: estimación de los parámetros, información y otros conceptos 155

1. Introducción a la estimación de los parámetros de los modelos 155

1.1. Estimación de la aptitud θ de los sujetos 156

1.2. Las puntuaciones estimadas en la TRI 161
2. Calibración de los ítems o estimación de los parámetros del ítem 162
3. Calibración de los ítems de un test y estimación de las puntuaciones de los sujetos: ejemplo de cálculo con un test real 163
4. Cumplimiento de los supuestos y ajuste de los datos al modelo de TRI elegido .. 167
4.1. Prueba de los supuestos de los modelos 167
4.2. El ajuste de los modelos a los datos 170
5. La escala de aptitud θ y la curva característica del test 173
6. La información de los ítems y del test 177
6.1. La información del ítem 177
7. La información del test $I(\theta)$ 181
7.1. La función de información del test 181
7.2. La función de información del test y el error típico de estimación 184
7.3. Estimación de la aptitud, información de ítems y test y errores típicos de estimación con el programa BILOG-MG 186
7.4. La función de eficiencia relativa y su papel en la construcción de un test .. 189

7. Modelos Politómicos de la Teoría de la Respuesta al Ítem 193
1. Introducción 193
2. Modelo de Respuesta Graduada (MRG) 195
3. Modelo de *Rating Scale* (Muraki, 1990) 203
4. Modelo de Crédito Parcial (MCP) 208
5. Modelo de Crédito Parcial Generalizado (MCPG) 212
6. *Software* para la estimación de los modelos de la TRI 216

8. Validez de los tests 219
1. Introducción 219
2. La evolución del concepto de validez 221
3. Una visión general del proceso de validación de tests 226
4. Argumentos de validez y sus fuentes de evidencia 229
5. Evidencias de validez basadas en el contenido del test 229
5.1. Concepto 229
5.2. Procedimientos para el análisis de las evidencias de validez de contenido 231
6. Evidencias basadas en el análisis de los procesos de respuesta 233
7. Evidencias basadas en la estructura interna (análisis psicométrico interno del test) 234
8. Evidencias basadas en las relaciones con otras variables 235
8.1. Las evidencias convergente y discriminante 236
8.2. Las evidencias referidas a criterios 240
9. Evidencias basadas en las consecuencias del test 243
10. El tratamiento de las evidencias de validez en las investigaciones psicológicas 244

9. Evidencias de validez referidas a un criterio, I: una variable predictora 245

1. El coeficiente de validez como correlación entre un test y un criterio 245

1.1. Introducción 245
1.2. Las ecuaciones de regresión lineal y el criterio de mínimos cuadrados ordinarios 246
1.3. Conceptos utilizados en la Teoría de la Validez 248
1.4. Inferencias en el modelo de regresión 251

2. Factores que afectan al coeficiente de validez como correlación entre puntuaciones del test y del criterio 258

2.1. Relaciones entre fiabilidad y validez de las puntuaciones de *X* y de *Y* 259
2.2. Cambios en el coeficiente de validez cuando se producen cambios en la longitud de los tests 263
2.3. Cambios en el coeficiente de validez con cambios en la variabilidad del grupo: el problema de la restricción del rango 265

3. Generalización de la validez 276

10. Evidencias de validez referidas a un criterio, II: validez y utilidad 279

1. Decisiones derivadas del uso de los tests 279
2. Utilidad de las decisiones en selección 281
3. Decisiones con los tests de diagnóstico clínico 284

3.1. Algunos índices utilizados en la validación de tests 284
3.2. El análisis mediante las curvas ROC 288

11. Evidencias de validez referidas a un criterio, III: múltiples predictores. Correlación y Regresión Lineal Múltiple 295

1. Decisiones derivadas del uso de los test 295
2. Modelo de Regresión Lineal Múltiple (RLM) 296

2.1. Supuestos 297
2.2. Estimación de los parámetros del modelo 297
2.3. Coeficientes de la regresión 298
2.4. Descomposición de la variación total observada en *Y* 299

3. Coeficientes de determinación múltiple y de correlación múltiple 299
4. Ejemplo de un modelo de regresión lineal múltiple 301
5. Inferencias acerca del modelo ajustado 303

5.1. Significación global del modelo 303
5.2. Significación estadística de los coeficientes de la regresión 304
5.3. Una interpretación simple de los valores de los coeficientes 305

6. Valores pronosticados, residuos e intervalos de confianza 306
7. Selección de predictores 307

7.1. La multicolinealidad en los predictores 308
7.2. Significación estadística de un predictor añadido 309
7.3. Contribución de los predictores al pronóstico 310
7.4 . Algo más sobre los procedimientos paso a paso 312

12. Evidencias de validez relativas a la estructura interna del test, I: Análisis Factorial Exploratorio (AFE) 317

1. Introducción 317

1.1. Introducción 317
1.2. Aproximaciones actuales al AF 319
1.3. Ejemplo para las ilustraciones de las técnicas 321

2. Conceptos básicos del Análisis Factorial Exploratorio 322

2.1. Fases del Análisis Factorial Exploratorio 322
2.2. Modelo del AFE y la ecuación fundamental 324
2.3. Algunos conceptos básicos del modelo de AF 325

3. Extracción de los factores 327

3.1. Métodos de extracción de los factores 330
3.2. Número de factores a extraer 333

4. Soluciones factoriales indirectas: la rotación 337

4.1. Algunas matizaciones de las soluciones oblicuas 341
4.2. Los algoritmos de rotación de los programas de ordenador 341

5. Interpretación de los factores 343

5.1. Tamaños muestrales requeridos para realizar Análisis Factorial y la replicabilidad de los factores 344

6. Factores de orden superior 344
7. Puntuaciones factoriales 345
8. Consideraciones en el Análisis Factorial de ítems de los tests 347

8.1. Problemas en la aplicación del AF a ítems de tests 347
8.2. El caso particular de los ítems dicotómicos 349

13. Evidencias de validez relativas a la estructura interna del test, II: Análisis Factorial Confirmatorio 353

1. Análisis Factorial Confirmatorio 353
2. Especificación del modelo 355

2.1. La matriz de varianzas-covarianzas del modelo 357

3. Otras consideraciones previas a la estimación del modelo 357

3.1. Preprocesamiento de los datos 357
3.2. Elección de la matriz para el análisis 358

4. Identificación del modelo 359
5. Estimación de los parámetros del modelo 361

5.1. Mínimos Cuadrados no Ponderados (ULS) 362
5.2. Mínimos Cuadrados Generalizados (GLS) 362
5.3. Máxima Verosimilitud (ML) 363
5.4. Estimación de parámetros en el modelo del ejemplo 363

6. Evaluación del ajuste de los modelos 365

6.1. Medidas globales de bondad de ajuste 366

6.2. Estadísticos de ajuste global para los datos del ejemplo 371
6.3. Ajuste de los componentes del modelo 371

7. Reespecificación de los modelos 373
8. Comparación de modelos anidados 377

14. Equidad del test y funcionamiento diferencial de los ítems 379

1. Introducción 379

1.1. Concepto de equidad o ausencia de sesgo 379
1.2. Preocupación histórica por el estudio del sesgo 381

2. Detección del sesgo 382
3. Procedimientos empíricos para la detección del DIF en ítems dicotómicos 385

3.1. Procedimientos no paramétricos 385
3.2. Procedimientos paramétricos: regresión logística 391

4. Procedimientos basados en estimaciones de la variable latente para la detección del DIF en ítems dicotómicos: métodos basados en Teoría de la Respuesta al Ítem 394

4.1. Definición y tipos de DIF en la TRI 394
4.2. Procedimientos paramétricos 396
4.3. Procedimientos no paramétricos basados en la variable latente 399

5. Modelos basados en tablas de contingencia frente a modelos basados en la TRI .. 399
6. Procedimientos de detección del DIF para ítems politómicos 400

6.1. Dificultades especiales de análisis del DIF en ítems politómicos 400
6.2. Procedimientos paramétricos basados en variables latentes 402

7. Algunas consideraciones finales 402

15. Interpretación de las puntuaciones: escalas, normas y equivalencias de puntuaciones 405

1. Introducción 405
2. Interpretaciones normativas 406

2.1. Introducción 406
2.2. Transformaciones lineales y no lineales 407
2.3. Otros tipos de normas 416
2.4. Grupos normativos 416

3. Equivalencia de puntuaciones y técnicas para lograr la comparabilidad 420

3.1. Introducción 420
3.2. Condiciones para la equiparación 421
3.3. Diferentes formas de lograr la comparabilidad 423
3.4. Diseños para la recogida de datos 425

4. Algunos métodos de equiparación basados en la TCT 428

4.1. Equiparación en puntuaciones observadas en la TCT con grupos aleatorios 429
4.2. Métodos de equiparación para el diseño de grupos no equivalentes 434

5. Procedimientos de equiparación basados en la TRI 436
5.1. Calibraciones separadas de *X* e *Y* 436
5.2. Calibración concurrente con el programa BILOG-MG (PARSCALE, ICL). 437
6. Observaciones finales 438

16. Interpretación de las puntuaciones basada en criterios: Test Referidos a Criterios (TRC) y estándares de rendimiento 439
1. Introducción 439
2. Definición y principales características de los TRC 440
3. Principales diferencias entre TRC y Tests Referidos a Normas (TRN) 441
4. Construcción de un TRC 442
5. Análisis de los ítems en los TRC 444
5.1. Análisis de contenido de los items 445
5.2. Análisis estadístico de los items 446
6. Fiabilidad de los TRC 451
6.1. La fiabilidad de los TRC en las puntuaciones del dominio 451
6.2. La fiabilidad de las clasificaciones 452
7. Determinación de puntos de corte para establecer los niveles 457
7.1. Procedimientos de establecimiento de los puntos de corte 459

Bibliografía 465

Prólogo

Este libro es el resultado de nuestra experiencia en la enseñanza de la asignatura de Psicometría a los estudiantes de Psicología de la Universidad Complutense de Madrid y de una extensa práctica en la construcción de instrumentos de medida. El objetivo fundamental de la asignatura que impartimos es introducir a los estudiantes en los principales conceptos, métodos y procedimientos de cálculo para construir, evaluar y aplicar tests en psicología, educación y en las ciencias sociales en general.

Desde nuestra visión, la Psicometría, como fundamento para la construcción y uso de los tests, debería estar contextualizada y guiada por conceptos sustantivos y no solamente como un mero conjunto de ejercicios estadísticos. Por este motivo, la mayor parte de los capítulos están ilustrados con conjuntos de datos reales similares a los que surgen habitualmente en la investigación psicológica.

Escribir el libro no ha sido fácil, ya que la Psicometría constituye hoy un extenso cuerpo de conocimientos, basado en diferentes modelos estadísticos. Algunos, de una gran tradición y antigüedad como la Teoría Clásica de los Tests, han proporcionado los fundamentos para el desarrollo de la mayor parte de los tests que los psicólogos utilizan en la actualidad, sean éstos de aptitudes, personalidad o rendimientos académicos. En los últimos treinta años se han ido incorporando con fuerza otros modelos como la Teoría de la Generalizabilidad y la Teoría de la Respuesta al Ítem, que juegan hoy un importante papel en la medición psicológica y educativa, en lo que se refiere al tratamiento del error y de la fiabilidad. Los siete primeros capítulos del libro explican los conceptos básicos de los tests, así como los fundamentos de las citadas teorías, imprescindibles para explicar las puntuaciones.

En la teoría de los tests no solamente son importantes los modelos para las puntuaciones, sino que aún lo es más el concepto de validez que permite fundamentar las inferencias y usos derivados de las puntuaciones. Aunque insuficientemente tratado en muchos textos de Psicometría, la validez es el aspecto más importante de un test y debe ser considerado desde muchos puntos de vista, dependiendo del uso que se pretende dar a los datos. Esto lleva a una gran variedad de procedimientos estadísticos, que quedan recogidos en el libro en los capítulos 8 a 13.

Los psicólogos y otros usuarios de los tests deben utilizarlos de forma justa y equitativa, no discriminando con su uso a determinadas poblaciones y el capítulo 14 ofrece una visión de esta problemática y de algunos procedimientos estadísticos para abordarla.

Finalmente, en los capítulos 15 y 16 se presentan las dos formas de interpretación de los tests, conocidas en la literatura como tests referidos a normas y tests referidos a criterios.

La cobertura de temas es extensa, lo que implica que no puedan ser tratados en profundidad, ya que serían necesarios libros monográficos para la mayor parte de los tópicos considerados. El libro intenta desarrollar una visión aplicada de la psicometría que resulte comprensible para los lectores. Por este motivo se ha prescindido de las demostraciones formales, insistiendo más en las aplicaciones y en el uso de *software* para la realización de los cálculos y en la interpretación del *output* de los distintos programas. En todos los casos se remite a los lectores interesados a numerosas referencias en las que pueden encontrar justificaciones rigurosas.

Una advertencia importante para los lectores se refiere a la notación utilizada en los capítulos. La psicometría procede de muchas fuentes teóricas y de muy diversos autores y grupos, que desarrollan notaciones específicas y que no tienen en cuenta las de otras fuentes. Esta diversidad de notación se aprecia sobre todo en los temas de modelos de puntuaciones/fiabilidad y los de validez. En el primer caso, en las ecuaciones suelen utilizarse letras griegas, mientras que en los temas referidos a la validez es común la utilización de letras latinas. En general, se ha seguido este tipo de notación, respetando la que es habitual en la mayor parte de los textos. Esperamos que no cree confusión a los lectores.

Con el objetivo de no alargar excesivamente el libro y no perder continuidad en su lectura, no se incluyen ejercicios prácticos ni instrucciones para la aplicación de los programas utilizados en los ejemplos que se presentan. No obstante, somos conscientes de la extraordinaria utilidad de los ejercicios prácticos y por eso se incluyen en el acceso al título de este manual, *Psicometría*, a través de la página web de la editorial (www.alianzaeditorial.es), una serie de ejercicios, tutoriales y datos con los que se pueden realizar prácticas.

LAS AUTORAS
Madrid, julio de 2006

1. Tests psicológicos y educativos: conceptos básicos, clasificación y construcción

1. Introducción

1.1. La medida en psicología y educación

La meta de las ciencias es el establecimiento de principios generales que permitan la descripción, predicción y explicación de los fenómenos de su área de interés. Para lograr este objetivo, como señala Torgerson (1958), los científicos deben recoger y comparar datos que permitan fundamentar las teorías y la medida es un componente fundamental en este proceso. La especialidad de la psicología que estudia los procesos de medición es la psicometría, que es un término genérico que aglutina todo un conjunto de modelos formales que establece las bases para que la medición de los fenómenos psicológicos y educativos se realice de forma adecuada.

En la teoría de los tests las definiciones de medida son sencillas. Existe bastante acuerdo en considerarla como la asignación de números a atributos de los sujetos, de tal forma que los números reflejen los diferentes grados del atributo que es evaluado (Nunnally y Bernstein, 1994, De Vellis, 2003). Una definición similar es la dada por Lord y Novick (1968) que la definen como «un procedimiento para la asignación de números (puntuaciones) a propiedades especificadas de las unidades experimentales, de modo que las caractericen y preserven las relaciones especificadas en el dominio comportamental» (p. 17).

Algunos lectores considerarán un poco simplista la definición de medida seguida, ya que se han omitido aspectos importantes de las *teorías de la medición*,

de gran relevancia en la psicología. Estas teorías intentan vertebrar la medición psicológica en torno a tres grandes áreas o problemas: la representación, la unicidad y la significación. Los lectores interesados en esta problemática pueden acercarse a las obras de Krantz, Luce, Suppes y Tversky (1970), Luce, Krantz, Suppes y Tversky (1990), Suppes, Krantz, Luce y Tversky (1989) y a la interesante crítica de Michell (1999). El mundo de los tests ha estado bastante alejado de esta problemática, ya que se parte del supuesto de que los valores asignados a las respuestas forman una escala numérica de intervalo, aunque la mayor parte de las veces no se soportan las propiedades de dicha escala (Lord y Novick, 1968). Este tipo de medición apriorística, o por *especificación*, ha recibido diversos nombres en la teoría de la medición: medidas por definición (Torgerson, 1958), medidas de indicador (Suppes y Zinnes, 1963) y medidas de índice (Dawes, 1972). Estas diferentes denominaciones suponen distintas formas de caracterizar el estatus métrico de las puntuaciones de los tests con respecto a la teoría de la medición, que puede resumirse en una simple frase: «ausencia de la justificación de las propiedades de la escala» (esta característica no es aplicable a un tipo de modelos que serán tratados más adelante, los modelos de Rasch). Para justificar esta aproximación, Lord y Novick (1968) insisten en la necesidad de evaluar estas puntuaciones desde el punto de vista pragmático, es decir, desde la utilidad de la escala resultante, destacando que si las puntuaciones de los tests proporcionan más información para la clasificación y predicción cuando son tratadas como de intervalo, pueden usarse así.

A pesar de las críticas, a veces recibidas de los ortodoxos de la teoría de la medición, es un hecho indudable que los tests psicológicos y educativos están entre las contribuciones más importantes de las ciencias del comportamiento a la sociedad, proporcionando mejoras fundamentales sobre otros procedimientos de medición.

La medida mediante tests en psicología tiene una larga historia, en la que por limitaciones de espacio no podemos entrar. Recomendamos a los lectores interesados la lectura de Du Bois (1970).

1.2. ¿Por qué los profesionales necesitan conocer la teoría de los tests?

Cuando una persona se enfrenta a esta materia por primera vez, como sucede con los estudiantes de psicología o educación, suelen plantearse ciertos interrogantes de los que son ejemplo los señalados por Crocker y Algina (1986):

1. ¿Qué es la teoría de los tests y por qué su conocimiento es imprescindible en psicología y educación?
2. ¿Por qué se requiere para su estudio conocer la estadística?

Tests, escalas de calificación, autoinformes, cuestionarios y otros instrumentos similares forman parte de la actividad cotidiana de psicólogos y edu-

cadores, y están llegando al público en general. Diferentes características de los sujetos, las escuelas, los estados, etc., son evaluadas periódicamente e incluyen tests de desarrollo, capacidades académicas, aptitudes, rendimientos escolares, inteligencia, actitudes, motivación, etc. Por otra parte, los resultados de los tests suelen tener con frecuencia impacto en la vida de los sujetos u otras unidades medidas. Es necesario conocer los procedimientos en los que se basan las puntuaciones de los tests, la información que proporcionan y las limitaciones que entrañan, para hacer un buen uso de los mismos. Como señalan los *Standards for Educational and Psychological Tests* (AERA, APA, NCME, 1999, en adelante SEPT), «el uso inadecuado de los tests puede causar considerables perjuicios a los sujetos que contestan al test y a otras partes afectadas por las decisiones derivadas de los tests*»* (p. 1).

Puesto que los constructos medidos en psicología y educación son abstracciones que únicamente se pueden evaluar indirectamente, el diseño de instrumentos adecuados de medición presenta múltiples problemas que se pueden resumir en los siguientes:

1. No existe una única aproximación a la medición de un constructo que sea universalmente aceptada. Diferentes procedimientos de medición, que se derivan de distintas definiciones del marco, pueden llevar a distintas conclusiones sobre la caracterización de los sujetos en el atributo o rasgo latente. El término constructo es utilizado aquí en sentido amplio, como el concepto o característica que el test pretende medir (SEPT, 1999).
2. Todos los tests están basados en muestras limitadas de conductas del dominio que se pretende medir. Determinar el número de elementos de la muestra y su variedad es uno de los principales problemas en el desarrollo de los instrumentos de medida.
3. Las medidas obtenidas siempre contienen error. Puesto que se basan en muestras limitadas de contenidos, se ven afectadas por todos los errores de muestreo. Las inconsistencias en las puntuaciones ligadas al muestreo de tareas, ocasiones o situaciones pueden considerarse como errores aleatorios.
4. Falta de escalas de medida con origen y unidades de escala definidas, lo que da lugar a una indeterminación de la medida.
5. Los constructos no pueden definirse aisladamente, simplemente en términos de definiciones operativas, sino que deben establecerse sus relaciones con otros constructos y con las conductas observables.

En cuanto a sus relaciones con la estadística y la Teoría de la Probabilidad, estas disciplinas constituyen las bases teóricas que fundamentan la teoría de los tests, como el lector podrá comprobar en la lectura de los siguientes capítulos.

En la evaluación psicológica y educativa observamos lo que las personas dicen o hacen en circunstancias particulares e intentamos inferir lo que cono-

cen o son capaces de hacer. Existe una cadena de inferencias que dependen de modelos estadísticos y probabilísticos, y es con estos enlaces con los que se asocian las propiedades psicométricas a las que se refieren términos como fiabilidad, validez y comparabilidad. Las fórmulas y procedimientos de la teoría de los tests proporcionan definiciones de trabajo e instrumentos prácticos para tratar estos aspectos, que cumplen su función en la cadena de inferencias de las observaciones al constructo (véase más adelante la figura 1.1). Como señala Messick (1994), «validez, fiabilidad, comparabilidad y equidad, no son solamente cuestiones de medida, sino valores sociales que tienen significado y fuerza más allá de la medida siempre que se emiten juicios y se toman decisiones» (p. 2).

2. Concepto de test. Algunas clasificaciones

La palabra inglesa *test* tiene varios significados. Como nombre significa prueba, reactivo, etc. Como verbo, ensayar, comprobar, etc. Este término ha sido adoptado internacionalmente para designar un tipo de examen o evaluación de uso extendido en psicología y educación. Aunque se han propuesto muchas definiciones de tests (véase Martínez Arias, 1995, para una revisión de las mismas), adoptamos la definición presentada en los últimos SEPT (1999) que recoge muy bien la variedad de aspectos a los que se puede aplicar esta denominación: «Un test es un instrumento evaluativo o procedimiento en el que se obtiene una muestra de la conducta de los examinados en un dominio especificado y posteriormente es evaluada y puntuada usando un procedimiento estandarizado» (p. 3).

En algunos contextos a veces se establecen distinciones entre los términos tests, inventarios, cuestionarios o escalas. En las páginas que siguen utilizamos la palabra test para referirnos a cualquiera de los tipos de instrumentos de medición, tal como establecen los SEPT. Por lo tanto, en el desarrollo del proceso de evaluación, toda la teoría que explicamos puede aplicarse tanto a los tests de rendimiento, aptitudes, actitudes, personalidad, etc.

Aunque se han propuesto múltiples clasificaciones de los tests, presentamos aquí las más habituales en los contextos psicológicos y educativos:

1. En función de las *consecuencias* para el sujeto. Es frecuente dividir los tests o instrumentos de evaluación a lo largo de un continuo de consecuencias, aunque simplificadamente suele hablarse de tests de altas consecuencias (como, por ejemplo, tests utilizados en procesos de selección) y tests de bajas consecuencias (como, por ejemplo, cuando se utilizan en investigación). Esta distinción es una consideración importante en la interpretación de los resultados.
2. En función del *planteamiento del problema y tipo de respuesta*. Los tests están construidos a partir de una muestra de conductas que exi-

ge, por parte del sujeto, una determinada forma de planteamiento y resolución del problema presentado. Como veremos en capítulos siguientes, esta distinción suele plantearse en términos del denominado formato de los ítems o tareas del test, y en este ámbito es frecuente hablar de tests de respuesta seleccionada (RS) o elección múltiple (EM) y test de respuesta construida (RC), incluyendo ambos diferentes modalidades.

3. En función del *área del comportamiento acotada*. Se establecen distinciones frecuentes entre tests cognitivos (aptitudes, inteligencia, rendimiento académico, etc.) y no cognitivos (personalidad, intereses, motivación, actitudes, etc.), a veces conocidos como tests de ejecución o rendimiento máximo frente a tests de ejecución típica.
4. En función *de la modalidad de aplicación*. Con respecto a este aspecto suelen utilizarse diferentes etiquetas para los tests, tales como individual/colectivo, papel y lápiz/ordenador, adaptativo/no adaptativo, etc.
5. En función de *las demandas temporales*. Este aspecto permite clasificar los tests en un continuo que va de rapidez o velocidad a potencia.
6. En función del *grado de aculturación o demandas específicas* de una cultura o grupo requeridas en la resolución del test. No se establece una división dicotómica, sino que, como alguna de las dimensiones anteriores, permite situar los tests a lo largo de un continuo.
7. En función del *modelo estadístico* en que se basan las puntuaciones del test. La construcción de un test necesita basarse en algún modelo probabilístico que permita dar un significado a las puntuaciones y hacer inferencias a partir de la muestra de conductas planteadas en el test. Aunque se han presentado muchos modelos a lo largo de la historia de la psicometría, los que han tenido trascendencia pueden clasificarse en los dos grupos siguientes:

    a) Teoría Clásica de los Tests (TCT) y sus extensiones, como la Teoría de la Generalizabilidad (TG), a veces conocidos como Modelos Débiles de la puntuación verdadera, basados en el modelo lineal general y en las puntuaciones totales de los tests.
    b) Teoría de la Respuesta al Ítem (TRI) que modeliza las probabilidades de respuesta a un ítem para diferentes niveles del rasgo latente mediante un modelo no lineal.

8. En función del *tipo de interpretación de las puntuaciones.* Se establece una distinción entre:

    a) La que pone el acento en el nivel o grado de ejecución sobre algún dominio o criterio definido de antemano, que da lugar a los normalmente conocidos como Tests Referidos a Criterio, TRC (Criterion Referenced Tests).

b) La aproximación basada en normas, que sitúa a los sujetos o estímulos en función de estadísticos calculados en algún grupo de referencia al que pertenece el sujeto, que da lugar a los conocidos como Tests Referidos a Normas (TRN).

Otra diferenciación importante, que se tratará más adelante, es la que se establece en función de los usos pretendidos del test.

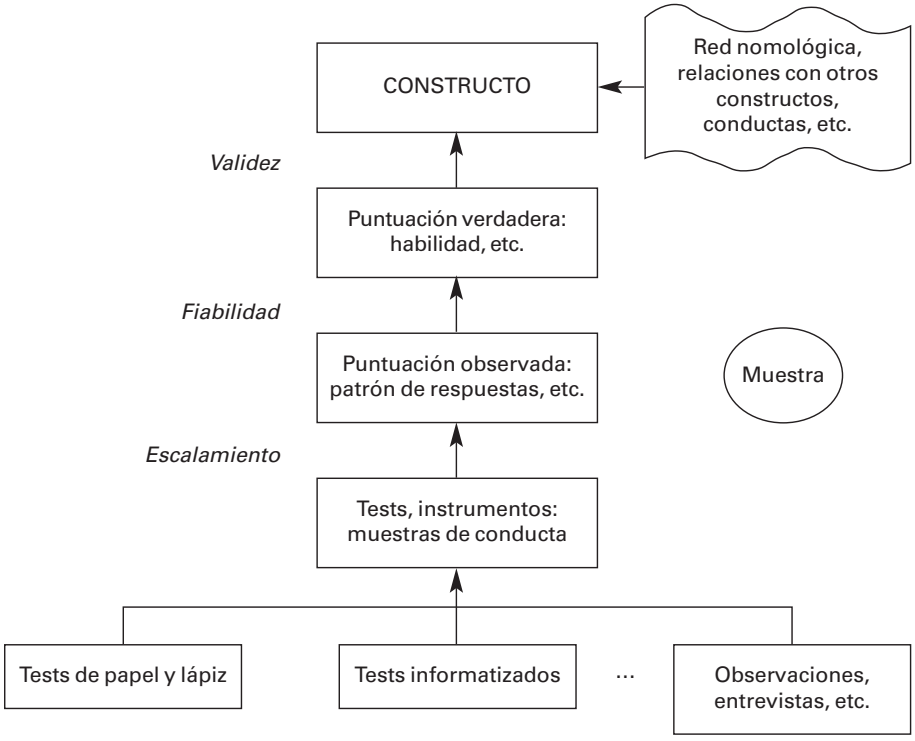
3. El proceso de inferencia psicométrica

Una dificultad a la que se enfrenta la psicometría es que la mayor parte de los atributos de interés en psicología y educación no son observables directamente, debiendo acudir a indicadores indirectos de los mismos: las conductas observables. Las teorías positivistas marcaron el lenguaje de la psicometría, refiriéndose a estas características no observables directamente como *constructos teóricos*, conectados por medio de las teorías con las conductas observables y con otros constructos. La relación de los constructos latentes con los hechos observables permite asignarles números y establecer relaciones empíricas entre ellos. Para la medida de estos constructos latentes se requiere algún instrumento que ponga de relieve los niveles variables del constructo y de algún modelo que relacione lo que se observa en el mundo real (las conductas observables) con el atributo o constructo que se pretende evaluar, que existe como parte de una teoría. Estas conductas observables de los sujetos suelen recogerse en un instrumento estandarizado, que suele denominarse test y como señala Wilson (2005), el propósito central de la medida en psicología y educación es proporcionar una forma razonable y consistente de resumir las respuestas de los sujetos a estos instrumentos en algún tipo de puntuación, que suele utilizarse para tomar decisiones sobre sujetos o grupos de sujetos. La idea central que subyace al uso del test es que hay un atributo único, el constructo, que es lo que el instrumento intenta medir y que en ese atributo, utilizando las puntuaciones del test, pueden situarse las personas y los estímulos o ítems del test.

Los constructos que intenta medir el test pueden ser muy variados: parte de un modelo cognitivo, la actitud de las personas hacia algo o un trastorno psicológico, pero en todos los casos, tras la aplicación del instrumento y siguiendo una regla que se aplica a las respuestas u observaciones, obtenemos una o más puntuaciones numéricas del sujeto. Suponemos que estas cantidades observadas reflejan de algún modo la cantidad que el sujeto posee en el atributo o propiedad medida, pero no podemos asumir directamente que refleje bien la verdadera cantidad de interés. El azar, en forma de errores aleatorios, juega un importante papel en estas mediciones o asignaciones numéricas y el proceso de medición se convierte en un proceso inferencial.

Esta problemática del proceso de inferencia psicométrica queda reflejada en la figura 1.1, a partir de la que examinaremos la complejidad de dicho proceso.

Figura 1.1. El proceso de inferencia psicométrica



Como puede observarse en la figura, en primer lugar se construyen una serie de tareas (los ítems del test) a las que se pide al sujeto que responda. Normalmente estas tareas se construyen a partir de las especificaciones del constructo o dominio de contenidos que conviene que sean lo más detalladas posibles. La presentación y las respuestas a estas tareas pueden adoptar múltiples modalidades, como puede observarse en la figura. Son las respuestas a estas tareas la única parte observable del proceso. Estas respuestas reciben una puntuación numérica que suele denominarse también puntuación observada por medio del proceso conocido como *escalamiento*. Existen diversos procedimientos según el modelo de puntuaciones seguido en la construcción del test, pero el resultado del proceso siempre es una puntuación, a veces denominada estimador muestral. Cuando aplicamos un test, pensemos por ejemplo en un test de habilidades matemáticas de 6.º de educación primaria, solamente se presenta una muestra de toda la población o dominio y la puntuación del sujeto solamente refleja la puntuación en un conjunto particular de tareas, que puede reflejar mejor o peor

la auténtica capacidad del sujeto y puede considerarse como un estimador muestral de ésta. El grado en que la puntuación observada representa la puntuación verdadera o su posición en el dominio (*V* en la TCT, puntuación del universo en la TG y θ en la TRI) se conoce como la fiabilidad del instrumento o precisión de la medida (en TCT), generalizabilidad o grado de confianza (en la TG) e información (en la TRI). A mayor fiabilidad, mayor seguridad en la medida, es decir, mayor correspondencia entre ambas puntuaciones.

Ahora bien, el test se ha construido con una finalidad u objetivo de medida, normalmente definido a través de un constructo o dominio y entonces entramos en el terreno de la *validez* con todas sus consideraciones de inferencias realizadas a partir de las puntuaciones de los tests.

Para realizar de modo adecuado la anterior cadena de inferencias, los autores y usuarios de los tests deberán apoyarse en la teoría psicométrica. Los constructos psicológicos se conceptualizan normalmente como variables latentes o no observables que subyacen a la conducta y que son las que influyen o determinan las respuestas o variables observables, tales como las puntuaciones de los tests o las respuestas a los ítems. Las respuestas de los sujetos a un ítem o a un test son un indicador de la posición de una persona en el rasgo latente, pero no define completamente la variable latente. Las medidas de constructos psicológicos son normalmente indirectas. Las variables latentes se miden observando conductas o tareas relevantes o ítems. Las propiedades de ambos, personas e ítems, sobre la dimensión latente se infieren de la conducta observable. La teoría psicométrica debe proporcionar las bases para relacionar las conductas con el constructo.

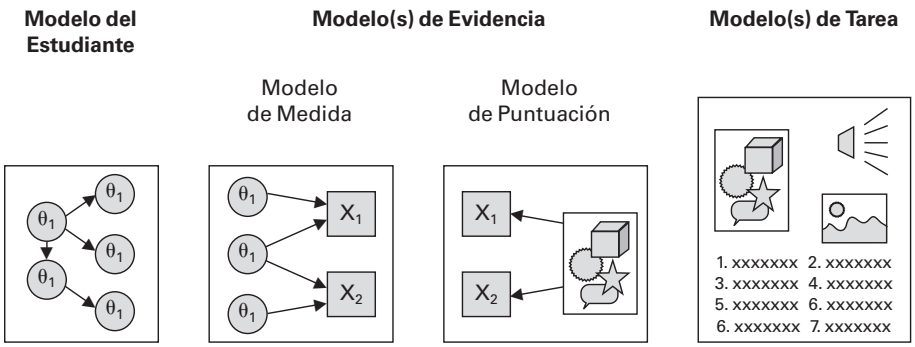
En la figura 1.1 las flechas de la cadena de inferencias están dirigidas de la tarea hacia el constructo, pero es el constructo el iniciador de todo el proceso, ya que es la teoría del constructo la que establece la selección de las tareas.

La anterior cadena de inferencias puede complementarse con lo que Mislevy, Wilson, Ercikan y Chudowsky (2003) denominan la perspectiva del *razonamiento evidencial*, que se refiere a cómo extraer inferencias sobre lo que los estudiantes conocen, pueden hacer o comprenden y que se representa en la figura 1.2. Una aportación interesante de esta perspectiva es que el constructo se establece en términos del *modelo del estudiante*, que se refiere a lo que queremos evaluar acerca de lo que un estudiante conoce, comprende o puede hacer. Siguiendo una tradición en psicometría a este constructo lo denotamos como θ. Este término puede referirse a una única categoría de conocimiento, como uso de vocabulario, o algo más complejo, como una serie de estrategias que el estudiante debe poner en juego. En el último nivel de la figura, nivel inferior en el modelo, tenemos las tareas, que son las situaciones en las que observamos lo que el estudiante hace. Estas tareas también precisan de un modelo. El modelo de tareas debe estar dirigido por el constructo o, como señala Wilson (2005), por el «mapa del constructo». El modelo de las tareas dirigirá el proceso de asignación de las puntuaciones *X*.

Entre el modelo de tarea y el modelo del estudiante (constructo) se encuentra el modelo de la puntuación y el modelo de medida que establecen el enlace entre lo que observamos y el constructo. El *modelo de la puntuación* se refiere al procedimiento que permite asignar las respuestas de los sujetos a unas categorías diferenciadoras (*X*), que normalmente se establecen en términos de algún tipo de ordenación.

El modelo de medida, conocido normalmente como modelo psicométrico, es el que nos permite conectar las puntuaciones con el constructo objetivo. Este modelo psicométrico es el que establece una serie de principios que garantizarán la adecuación de las inferencias psicométricas. Los usuarios de los tests deben ser conscientes de que raramente podemos inferir con certeza, ya que el conocimiento en el que se basa la inferencia psicométrica suele ser incompleto, pudiendo llevar a más de una explicación.

Figura 1.2. Modelo general de un diseño de evaluación (adaptado de Mislevy, *et al.*, 2003)



4. Los principios del modelo psicométrico

Aunque pueden recibir diferentes nombres en diferentes textos, básicamente los principios psicométricos que garantizan la calidad de las medidas, pueden reducirse según Mislevy, *et al.* (2003) a los siguientes: validez, fiabilidad, comparabilidad y equidad o ausencia de sesgos discriminatorios. Phillips (1996), por su parte, habla de validez, generalizabilidad, comparabilidad y equidad.

Aunque serán desarrollados en los capítulos posteriores, exponemos brevemente los aspectos que abordan cada uno de ellos.

4.1. Fiabilidad

Este principio tiene que ver con los errores cometidos en el proceso de medición, por lo que responde al problema de hasta qué punto las cantidades observadas reflejan con precisión la puntuación verdadera (puntuación del uni-

verso o aptitud) de la persona. En la práctica, tal como señala Brennan (2001b), tiene que ver con el proceso de repetición o generalización de la medida, idea que ha sido predominante desde los trabajos iniciales de Spearman (1904).

4.2. Validez

Es el más importante de los principios y nos habla del grado en que el uso que pretendemos hacer de las puntuaciones de los tests está justificado. Supone examinar la red de creencias y teorías sobre las que se asientan los datos y probar su fuerza y credibilidad por medio de diversas fuentes de evidencia. Requiere debilitar las explicaciones alternativas sobre el comportamiento del sujeto ante los ítems y eliminarlas para reducir los errores inferenciales.

Los primeros trabajos sobre validez distinguían entre un número de variedades de validez, tales como contenido, predictiva, convergente, discriminante y de constructo. En los SEPT actuales, la validez es el tema predominante y se define como un concepto unitario. Los diferentes tipos de validez son considerados como diferentes tipos de evidencia para un único tipo de validez. Si hace falta etiquetarla, la etiqueta sería la de *validez de constructo*.

Embretson (1983) distinguió entre dos tipos de evidencia: los que se refieren a la *representación del constructo* y los de la *red nomológica*, en la que se establecen relaciones con otros constructos. Hasta hace poco, las justificaciones de validez eran sobre todo del segundo tipo, pero desde la revolución cognitiva, hay fuertes bases para argumentos en la representación del constructo.

4.3. Comparabilidad

En ocasiones la recogida de datos se realiza de forma distinta para diferentes sujetos o para los mismos sujetos en diferentes momentos. Diferentes condiciones de medida hacen surgir hipótesis alternativas cuando se realizan comparaciones entre sujetos, o cuando se comparan con estándares o con el progreso en el tiempo. El problema de la comparabilidad responde a cuestiones tales como: ¿puede haber diferencias sistemáticas en las conclusiones si observamos las respuestas del test A en vez de las del test B?, ¿en un TAI (test adaptativo informatizado) frente a uno de lápiz y papel?, ¿con un evaluador frente a otro? A la hora de hacer inferencias deben eliminarse estas posibles explicaciones alternativas.

4.4. Equidad

En general, bajo esta problemática nos referiremos a explicaciones alternativas de los resultados de la evaluación que pueden estar relacionadas con factores tales como diferencias sociodemográficas, la historia personal de cada

individuo, lenguaje, familiaridad con las representaciones, etc. Cuando el mismo instrumento de medida se aplica a muchos sujetos (un grupo), es preciso examinar el impacto de dichos factores.

Es un concepto que hasta cierto punto también se solapa con otros aspectos, estando relacionado, además, con perspectivas sociales, políticas y educativas, sobre los usos de la evaluación (Willingham y Cole, 1997). En realidad, aunque será tratada separadamente, como señalan Moss y Schutz (2001a), forma parte de la teoría general de la validez.

5. Los modelos de tests

Como ya se ha señalado, para relacionar conductas observadas o respuestas con los constructos se usan normalmente modelos (*modelo de medida,* recogido en la figura 1.1). En el caso de la medida psicológica o educativa, son modelos matemáticos en los que variables independientes, los constructos, se combinan numéricamente para predecir de forma óptima una variable dependiente, la respuesta del sujeto.

Seguramente uno de los aspectos más característicos de la psicometría es el uso de modelos estadísticos. Los modelos de medida proporcionan reglas formales explícitas de cómo integrar las diferentes piezas de información que pueden ser relevantes para una inferencia particular. La idea esencial es aproximar las relaciones sustantivas importantes en términos de relaciones entre variables en un modelo probabilístico. Esto lleva a una versión simplificada de la realidad, que nos da reglas para su interpretación. En el caso, por ejemplo, de los tests de aptitudes, establecen las relaciones entre dichas aptitudes de los sujetos y su conducta observable en el test. Es decir, establecen las relaciones entre las variables latentes θ y las variables observadas X. No sabemos exactamente qué hará un sujeto en una situación particular, pero para los sujetos con un valor dado de θ, habrá una distribución de probabilidad de valores posibles de X, es decir, $P(X \mid \theta)$. Un modelo de medida no intenta explicar cada detalle de los datos, solamente aproximar los patrones importantes.

En el modelo deben considerarse varios aspectos. En primer lugar, el modelo especifica la escala para las observaciones o variable dependiente, que puede ser una puntuación, las respuestas a un ítem o una matriz de relaciones entre ítems y respuestas. Además, especifica una o más variables de diseño que son las variables independientes. Junto a esto, especifica cómo se combinan las variables independientes para predecir la variable dependiente. El modelo puede especificar una simple combinación aditiva de variables que predicen la variable dependiente, como en el modelo de la regresión lineal. Los pesos para las variables de diseño o independientes son los parámetros del modelo. Los modelos a veces especifican relaciones más complejas. Por ejemplo, pueden combinarse interactivamente o seguir un determinado tipo de función (p. ej., la función de distribución normal o logística) para relacionar las variables independientes con las dependientes.

Durante muchos años la TCT ha sido el modelo básico usado en la construcción de los tests. No obstante, y como observamos en varias partes de este libro, la complejidad de los problemas de medida, especialmente en el marco de las Evaluaciones a Gran Escala (EGE), lleva a situaciones que la TCT no es capaz de resolver de una forma eficaz. La popularidad creciente de la TRI se basa en que proporciona justificación teórica para hacer muchas cosas que no son posibles en el marco de la TCT. Algunos de los problemas, como veremos, también tienen soluciones desde la TCT, pero menos eficaces.

6. La construcción de un test

Un buen instrumento de evaluación debe ser construido de una forma rigurosa si queremos que sea adecuado para los usos a los que está dirigido. En la práctica, el desarrollo de un test es una tarea laboriosa que requiere de la colaboración entre expertos en la materia o constructo a evaluar y expertos en medición. Los tests pueden desarrollarse en el marco de dos grandes modelos, conocidos como Teoría Clásica de los Tests (TCT) y Teoría de Respuesta al Ítem (TRI).

Aunque el uso particular marcará algunas especificidades en el instrumento, hay una serie de normas que si se siguen, llevarán a buen puerto la tarea de construcción del test. Existe un conjunto de guías de buenas prácticas sobre el desarrollo y construcción de tests, cuya lectura recomendamos antes de embarcarse en la construcción de un test, entre las que se encuentran: *Standards for Educational and Psychological Tests* (AERA, *et al.,* 1999), *Standards for Quality and Fairness* (Educational Testing Service, 2000); y *Technical Standards for IEA Studies* (Martin, Rust y Adams, 1999).

En el terreno de los test educativos, otro complemento para aprender más sobre la construcción de instrumentos son los textos recientes sobre medida en educación a los que remitimos a los lectores: Haladyna (2004), Popham (2000), Nitko (2001) y Linn y Gronlund (2000). En el ámbito de la psicología social y de la personalidad son interesantes los textos de Netemeyer, Bearden y Sharma (2003) y De Velis (2003) y en castellano el de Morales, Urosa y Blanco (2003).

Aunque en ningún caso pueden obviarse las lecturas señaladas, presentamos algunas consideraciones sobre el desarrollo de los tests en la práctica. Revisaremos el proceso de construcción del test a partir de las fases siguientes:

1. Identificación del propósito o finalidad del test.
2. Restricciones para la aplicación.
3. Marco del test o mapa del constructo.
4. Especificaciones derivadas del marco.
5. Especificación del formato de los ítems y reglas de puntuación.
6. Revisión de los ítems del test.
7. Estudio piloto.
8. Estudio de campo.
9. Construcción del manual del test y sus revisiones.

6.1. La finalidad del test

El desarrollo de un test comienza con la delimitación del objetivo del test y el constructo o dominio que se pretende medir. En este terreno, un primer paso es considerar cuáles son las poblaciones a las que está destinado y el tipo de decisiones que se tomarán con las puntuaciones obtenidas.

Con respecto a la primera cuestión deberán determinarse las edades de los sujetos (o los cursos en muchos de los tests de rendimiento), si el test se aplicará a todos los sujetos o a grupos seleccionados de alguna forma (tests para la población general o destinados a grupos especiales), requisitos de lenguaje que deben tener los sujetos (pueden establecerse también modificaciones o adaptaciones permisibles), si se necesita o no una motivación especial para realizar el test, etc.

Con respecto a los posibles usos de las puntuaciones, sin ánimo de ser exhaustivos, entre los tipos de decisiones más frecuentes encontramos:

a) Decisiones relativas al nivel de instrucción o habilidad. El objetivo consiste en determinar si los sujetos poseen dominio de ciertos conocimientos o destrezas predeterminados de antemano. Suelen ser decisiones propias de los llamados Test Referidos a Criterio.
b) Decisiones de diagnóstico. Tienen como objetivo la identificación de fallos o deficiencias en el comportamiento para emprender alguna intervención o tratamiento. Aunque las puntuaciones suelen compararse con datos normativos, este tipo de decisión también es propia de los Tests Referidos a Criterio.
c) Decisiones de selección. El objetivo en este caso es la selección de los sujetos más capaces en algún sentido o materia. Las puntuaciones de los sujetos en el test se usan para admitir o excluir, emplear o rechazar.
d) Decisiones de asignación. El objetivo es determinar en qué lugar debe colocarse una persona dentro de una jerarquía. Esta jerarquía puede ser de carácter profesional (qué puesto de trabajo corresponde a un determinado sujeto), de carácter formativo (dentro de un programa de formación qué nivel le correspondería), etc.
e) Decisiones de clasificación. El objetivo radica también en asignar a los sujetos a ciertas categorías pero, en este caso, las asignaciones son de *tipo* en lugar de nivel. Por ejemplo: ¿dentro de qué categoría diagnóstica clínica clasificamos a un sujeto?
f) Decisiones de consejo. El objetivo es orientar al sujeto con respecto a algún aspecto de su comportamiento futuro, abarcando una gran cantidad de decisiones que el sujeto toma con ayuda de un profesional de la psicología o la psicopedagogía.
g) De *screening* (detección rápida): tiene como finalidad la detección rápida de algún tipo de trastorno, deficiencia o anomalía. El objetivo es proveer una herramienta sencilla, de rápida aplicación que permita

detectar personas que potencialmente puedan presentar un determinado trastorno, deficiencia, etc., sirviendo como «criba» antes de la aplicación de otras pruebas de carácter más exhaustivo.

6.2. Restricciones en la aplicación

Además de por la finalidad de uso, el desarrollo de un test estará condicionado por restricciones en la aplicación. Cualquier test debe plantearse siempre dentro de los límites impuestos por la práctica y el constructor del test deberá ajustarse a ellos.

Una de las limitaciones más claras es la del tiempo de aplicación, ya que los sujetos solamente pueden disponer de un tiempo limitado para la realización del test. Existe una estrecha relación entre la longitud del test y la fiabilidad de las puntuaciones, como se observa en el capítulo correspondiente a la TCT. Será necesario tener en cuenta esta cuestión, guardando un equilibrio entre ambas.

Otra cuestión de carácter práctico, que también impone restricciones, tiene que ver con cómo será la modalidad de aplicación del test. Esto es, si será aplicado de forma individual, colectiva, informatizada, etc. Así, por ejemplo, en el caso de aplicación colectiva, se deberá tener especial cuidado en las instrucciones, que deberán ser de fácil comprensión para la población a la que van destinados. En el caso de aplicación individual habrá que tener en cuenta aspectos tales como el tipo de material que se podrá manipular, tipo de ayuda que puede dar el examinador, número de intentos permitidos, etc. Estas consideraciones se aplican, de igual forma, a los test informatizados.

También es preciso matizar los materiales a utilizar durante la realización del test, tales como calculadoras, diccionarios, materiales de laboratorio, etc. Normalmente hace falta ajustarse a otros tipos de restricciones, como que los contenidos del test no se vean afectados por factores irrelevantes para la medida del constructo, p. ej., que un test de matemáticas requiera destrezas lingüísticas en las que existen diferencias individuales entre los sujetos, que el material estimular del test favorece a los varones frente a las mujeres, etc.

6.3. Marco del test: características del constructo o dominio

A partir del propósito del test y del dominio o constructo que se pretende medir se desarrolla el marco sobre el que comenzará el diseño del test.

Una aproximación al instrumento de evaluación centrada en el constructo comenzará preguntándose qué complejo de conocimientos, destrezas u otros atributos serán evaluados, presumiblemente porque estén ligados explícita o implícitamente a objetivos de la instrucción o son valorados por la sociedad. A continuación, qué conductas o ac-

tuaciones revelarán estos constructos, y qué tareas o situaciones elicitarán estas conductas. De esta forma, la naturaleza del constructo guía la selección o construcción de tareas relevantes, así como el desarrollo racional de criterios y rúbricas de puntuación basadas en el constructo.

(Messick, 1994, p. 16)

La cita de Messick establece los ingredientes necesarios para comenzar con el desarrollo del test: 1) se parte del constructo que se pretende medir, que, por ejemplo, en el caso de la evaluación educativa las cuestiones del test deben corresponderse con metas importantes de los programas, el currículo y la instrucción; 2) se especificarán las conductas que revelarán esos constructos, y 3) qué tareas o situaciones elicitarán dichas conductas.

Como se ha señalado anteriormente, el fenómeno o constructo que una escala intenta reflejar se denomina variable latente. Se caracteriza porque no es directamente observable pero se plantea como causa de las respuestas a los ítems construidos para medirlo. Es una variable, es decir, su fuerza o magnitud cambia y esto es precisamente lo que la escala pretende estimar, la magnitud actual para cada persona, en el tiempo y lugar de la medida.

Una tarea de fundamental importancia para la construcción del test que ha de medir un determinado constructo es la definición precisa del mismo. El constructo ha de estar claramente definido, así como también lo han de estar sus facetas y dominio. Es el primer paso y, probablemente, el más difícil en el proceso del desarrollo de escalas de medida. Todo constructo está basado en un marco teórico, disponer de una teoría bien especificada, normalmente disponible en la bibliografía, se convierte en un elemento imprescindible. Una revisión exhaustiva de la bibliografía puede revelar intentos anteriores de medir el constructo, y las fuerzas y debilidades de esos intentos. Cada vez se insiste más en el establecimiento de modelos teóricos del constructo para el desarrollo del test. Solamente una representación cuidadosa del constructo evitará la omisión de aspectos importantes, esto es, una infrarrepresentación del constructo, o la inclusión de conductas irrelevantes para el constructo. El proceso de construcción de tests guiado por la teoría es el más recomendable y el único que llevará a tests con fuertes evidencias de validez de constructo. A pesar de las demandas en este sentido que llegan desde la psicología cognitiva, este aspecto está algo descuidado en los tests psicológicos y educativos actuales (Pellegrino, *et al.,* 2001).

Una vez delimitado el constructo, el equipo responsable tendrá que especificar las principales conductas observables representativas del constructo (dominio o atributo de interés) e identificar las tareas mentales o procesos que el sujeto debe realizar en la materia particular, tales como recuerdo, análisis, aplicación, descubrimiento, etc. Estos aspectos necesarios al comienzo del desarrollo del test, se corresponden con lo que en epígrafes anteriores denominábamos los modelos del sujeto (conductas o rendimientos que se quieren medir) y el modelo de tareas para elicitar dichas conductas. El modelo del

sujeto especificará las variables en términos de lo que intentamos que caracterice a los sujetos. Los modelos de tareas son esquemas para obtener datos que proporcionen evidencias sobre ellas.

En la elección de las conductas representativas del constructo, la mayor parte de las referencias se centran en tests de rendimiento académico, existiendo muy poco escrito en psicología sobre cómo la teoría sobre el constructo se plasma en elementos o ítems del test. Normalmente en psicología se parte de un modelo teórico en el que se inscribe el constructo y una serie de conductas observables ligadas a éste. Si el dominio teórico no está suficientemente especificado, suelen recomendarse actividades como: *análisis de contenidos* donde se plantean preguntas abiertas a los sujetos sobre el constructo de interés clasificándose posteriormente en categorías, *revisión de las investigaciones publicadas*, estudio de *incidentes críticos,* que consiste en la identificación de conductas que caracterizan a los extremos del constructo de interés, realización de *observaciones directas, entrevistas con informadores clave*, *grupos de discusión*, *juicios de expertos* o a través de los *objetivos instruccionales o de programas de intervención,* que especifican normalmente las conductas observables que los sujetos mostrarán tras la aplicación del programa.

En el caso de los test educativos, la elección de las conductas representativas suelen establecerse a partir del currículo o del consenso de los expertos. En este caso el proceso incluye consideraciones sobre cómo se usarán los resultados, las consecuencias (deseadas y no deseadas) del uso, los estándares de contenido articulados, el material cubierto por el currículo, y las formas en que los estudiantes mostrarán el dominio del material.

6.4. Especificaciones derivadas del marco

Una vez que se ha establecido el propósito de la evaluación y el marco, deben desarrollarse las especificaciones del test que describan el marco anterior con más detalle, identificando las conductas concretas, destrezas o conocimientos que se evaluarán, las proporciones de ítems para cada uno de los aspectos, etc. Es conveniente también incluir especificaciones estadísticas, referidas a aspectos como las dificultades de los ítems, valores de coeficientes de fiabilidad y errores típicos de medida, etc.

En la generación del conjunto de ítems es fundamental revisar todas las áreas de contenido del constructo realizando un muestreo representativo o con elementos proporcionales a las facetas. En la construcción de un test normalmente se parte de un conjunto de ítems mucho mayor que el que va a constituir la versión definitiva. Esto permite al constructor seleccionar aquellos ítems que se adaptan mejor a las especificaciones del test. A partir del conjunto inicial, la elección de los ítems para la versión definitiva tendrá lugar a través de procedimientos de revisión del test y el estudio piloto, que serán tratados en apartados siguientes.

6.5. Especificación del formato de los ítems. Reglas de puntuación

La primera cuestión hace referencia al formato de los ítems, aspecto esencial en el desarrollo de los instrumentos. ¿Serán diseñados para corrección automatizada, demandarán respuestas construidas orales o escritas, o emplearán una combinación de ambos procedimientos?

El propósito del test y el dominio o constructo a evaluar también impondrá ciertas restricciones al formato de los ítems. Así, en los test de inteligencia y aptitudes las alternativas están entre la «construcción» de la respuesta por parte del sujeto y la «selección» entre un conjunto de alternativas planteadas. Si bien se ha tendido al uso de esta última opción, porque permite plantear un mayor número de preguntas en el mismo período de tiempo y su rapidez y facilidad de puntuación, las críticas a este tipo de formato cuestionan su uso indiscriminado (Bennett y Ward, 1993). En los test de personalidad, sin embargo, la selección de la respuesta es el procedimiento habitual.

Siguiendo a Bennett (1993), entre los posibles formatos de respuesta, ordenados de menor a mayor elaboración requerida, encontramos:

1. Elección múltiple: el sujeto ha de elegir entre un conjunto limitado (entre 3 y 5) de respuestas posibles.
2. Selección o identificación: el sujeto ha de elegir entre un número de respuestas posibles pero, a diferencia del anterior, el número es grande, evitándose de esta forma el efecto del azar. Un ejemplo de este tipo de formato sería: «eliminar palabras redundantes de un texto».
3. Ordenación del material estimular. El sujeto ha de disponer los estímulos en la secuencia correcta, por ejemplo, ha de ordenar frases o dibujos para que sigan una secuencia lógica, expresiones matemáticas para que constituyan una demostración, etc.
4. Sustitución/corrección. El examinado ha de sustituir algún elemento por una alternativa correcta. Corregir errores gramaticales u ortográficos, corregir sentencias de un programa de ordenador, etc., serían ejemplos de este formato.
5. Completar. El sujeto ha de completar algún estímulo que se presenta de forma incompleta, como por ejemplo, dar una respuesta numérica a un programa de cálculo, indicar qué palabra completa una frase, etc.
6. Construcción. En este caso, el sujeto ha de generar la respuesta completa. Por ejemplo, hacer un dibujo, un ensayo, gráficos a partir de un conjunto de datos, etc.
7. Presentación. El sujeto es evaluado bajo condiciones reales o simuladas. Es muy utilizado en selección de personal, ejemplos de este tipo de formato son las «muestras de trabajo».

En caso de elegir el formato de «selección» de respuesta, la segunda decisión es el número de alternativas que se van a ofrecer. En los test de rendi-

miento, inteligencia y aptitudes variarán de verdadero-falso a formatos entre 3 y 6 alternativas y en los test de personalidad los formatos varían entre «dicotomías» (sí-no) y «varios grados» (normalmente entre 4 y 9). En este último caso, un elemento importante es el enunciado verbal de los puntos de referencia. Otros formatos como las escalas tipo Thurstone y Guttman (cuyos ítems varían en la intensidad del atributo) tienen escaso uso en la actualidad. Y formatos como las ordenaciones suelen crear dificultades desde el punto de vista psicométrico.

Dentro de las escalas graduadas con más amplio uso se encuentran las escalas tipo Likert y el diferencial semántico, que admiten una gradación de las respuestas. Las escalas tipo Likert constan de un enunciado seguido de opciones de respuesta que indican grados variables de acuerdo o de frecuencia. Los puntos de la escala deberían reflejar intervalos de la misma distancia. El diferencial semántico (Osgood y Tannenbaum, 1955) se usa para escalar estímulos: objetos, grupos de personas, etc. En cada ítem hay un par de adjetivos que representan los extremos de un continuo: honesto… deshonesto; tranquilo… ruidoso, etc. Otro tipo de escala graduada de uso también en la actualidad es el análogo visual. Presenta una línea continua entre un par de descriptores que representan los opuestos de un continuo. Entre sus virtudes se encuentra que evita el anclaje verbal, es potencialmente más sensible y elimina la posibilidad del recuerdo de las respuestas.

Otras consideraciones importantes sobre el formato son las siguientes:

1. Los formatos dicotómicos llevan a distribuciones de puntuaciones muy sesgadas, pero necesitan menos tiempo de respuesta. También reducen la varianza de la escala, siendo necesaria la inclusión de más ítems.
2. Los formatos graduados requieren decidir sobre el número de puntos, que oscila normalmente entre 3 y 9. Cuando los números llevan etiquetas verbales, es preciso tener en cuenta si éstas pueden ser lo suficientemente diferenciadoras.
3. El hecho de tener un número impar de puntos permite disponer de un punto medio o neutral, un número par por el contrario fuerza a una mayor discriminación.

Otra cuestión fundamental en el desarrollo del test es el de la escritura de los ítems. Con respecto a la cual conviene cuidar los siguientes aspectos:

1. Claridad de los enunciados. Los ítems han de ser redactados evitando la ambigüedad, de modo que todos los sujetos los comprendan de la misma forma. El lenguaje ha de ser adecuado a la población objetivo cuidando, además, el nivel de dificultad de la lectura. Los ítems han de ser cortos y sencillos, ya que una longitud extensa reduce la claridad. No son convenientes tampoco los comienzos negativos.

2. Evitar la redundancia. Es importante evitar las reiteraciones innecesarias del texto, ya que aumentan el tiempo de lectura del sujeto no aportando nada a su comprensión. Hay que señalar que una cierta redundancia de ítems es necesaria, no obstante, hay redundancias inútiles, por ejemplo el cambio en una palabra. Es importante cambiar las palabras y la estructura sintáctica.
3. Considerar si es conveniente la presencia de ítems positivos y negativos. Esto permite evitar sesgos de aquiescencia y afirmación. No obstante, hay que tener cuidado a la hora de los análisis de la escala porque en algunos casos esto puede contribuir a un factor de método. Conviene, en cualquier caso, no comenzar con «no» y cuidar la coherencia con la escala de respuesta.

6.6. Revisión de los ítems

El proceso de revisión de los instrumentos puede prevenir errores costosos, por lo que conviene revisar los materiales desde varias perspectivas. En la revisión del test participarán expertos en el contenido, sujetos de las poblaciones objetivo y expertos en psicometría. Los especialistas en contenido deben revisar los instrumentos asegurándose que no ocurran problemas de infrarrepresentación del constructo o introducción de varianza irrelevante al constructo. Expertos en medida deben revisarlos desde el punto de vista metodológico. También conviene que sean revisados por sujetos pertenecientes a la población objetivo con diferentes experiencias y posiciones, para evitar sesgos de los tests, evitar situaciones de lenguaje ofensivas a grupos particulares, etc.

6.7. Estudio piloto

El propósito de esta fase es probar el funcionamiento del test, recabando información relativa a los métodos de recogida de datos, estímulos, materiales y cuestiones, validez y fiabilidad, criterios de puntuación, etc., que permitirá mejorar los instrumentos para la versión definitiva. La prueba se realizará sobre una muestra de sujetos, que ha de ser de características similares a la población objetivo.

Es importante recoger información de los sujetos sobre si comprenden las instrucciones, las cuestiones y los procedimientos de respuesta. Para ello, es importante desarrollar un cuestionario para que los que hacen el estudio piloto recojan las incidencias y dificultades implicadas, pudiéndose realizar también entrevistas después de la sesión. En esta fase es importante la recogida de tiempos de realización, para ajustar los tiempos finales. Como hemos señalado, normalmente se cuenta con muchos más ítems de los que se usarán

en el test definitivo y en algunas aplicaciones, se introducen ítems «variantes» que no se utilizan en la puntuación, y que se pilotan para futuras aplicaciones.

En esta fase también se realizarán ya algunos análisis estadísticos de los ítems para ver su funcionamiento, pudiéndose obtener información sobre la dificultad de las cuestiones y las relaciones de los ítems con el instrumento. Los análisis de los ítems son muy importantes para poder seleccionar de entre todos ellos los que se adapten mejor al objetivo del test. En el capítulo 4, en el apartado de análisis de elementos, se desarrollará en profundidad esta cuestión. Es importante también examinar el rendimiento de diferentes subgrupos para evaluar problemas potenciales de sesgo. En el capítulo 12 se tratará este tema. Finalmente, el estudio piloto permitirá, igualmente, examinar las propiedades estadísticas de los tests completos.

6.8. Estudio de campo

En los tests referidos a normas, la puntuación de cada sujeto se interpreta en relación a los otros sujetos de la población a la que pertenece, por lo que necesitamos datos procedentes de grupos de referencia. Los grupos de referencia están constituidos por las poblaciones y subpoblaciones a las que va dirigido el test. Normalmente, en los test referidos a criterio no es necesario, siendo el rendimiento o conducta de una persona en términos absolutos suficiente para describir su estado con relación al dominio establecido.

El test será aplicado a una muestra representativa de la población, que recibe el nombre de grupo normativo, que servirá de grupo de referencia para interpretar las puntuaciones de los sujetos. Así pues, el primer paso consiste en la selección de la muestra representativa de la población o subpoblaciones a las que va dirigido el test. En caso de sospecha de que puede haber diferencias entre subpoblaciones, deben establecerse los estratos necesarios (p. ej., sexos, profesiones, edades…).

Una vez aplicado el test a los sujetos seleccionados se estimarán a partir de los datos los estadísticos del test: fiabilidad, errores de medida, errores de clasificación, validez para el uso al que está destinado, estructura interna, etc. En el caso de los test referidos a criterio construidos desde la TRI, la muestra servirá para la calibración del test (estimación de las propiedades de los ítems). A partir de los datos se obtendrán las escalas de puntuaciones en las que están basadas las normas. En el caso de tests referidos a normas, se obtendrán las normas o baremos. Si está basado en estándares o criterios habrá que determinar los puntos de corte que establecen los distintos niveles.

6.9. Manual del test y sus revisiones

La elaboración de un manual del test es imprescindible cuando el test ha sido construido para su comercialización o para que sea utilizado por otros profesionales distintos al constructor. En él deberán incluirse los siguientes aspectos:

1. Fundamentación teórica del constructo o explicación del dominio que cubre el test.
2. Usos a los que se puede destinar el test.
3. Poblaciones a las que está dirigido.
4. Instrucciones completas para su administración.
5. Coeficientes de fiabilidad o datos de precisión de la medida.
6. Evidencias de validez para los usos a los que está destinado.
7. Sugerencias sobre posibles usos de las puntuaciones de los tests.
8. Normas de intepretación de las puntuaciones.

Por último, hemos de señalar que cambios en el dominio medido mediante un test, nuevos datos científicos sobre el constructo a evaluar o nuevas condiciones de uso de un test son factores que podrían cuestionar la validez de un test en el estado o uso actual o sugerir modificaciones en el contenido del mismo. De ahí, la necesidad de que el test sea revisado periódicamente para ver la conveniencia de cambios. Otra cuestión no referida al contenido pero que también necesita revisión es la actualidad de las normas de interpretación en los test de carácter normativo.

7. Criterios para la valoración de un test

A la hora de valorar un test de cara a su posible utilización existen una serie de cuestiones que el potencial usuario del mismo debe plantearse.

Presentamos una posible guía de éstas siguiendo normas extraídas de los SEPT.

1. ¿Está bien definido el marco y las especificaciones del constructo o dominio?
2. ¿Existen instrucciones estandarizadas y adecuadas a la población a la que va dirigido, e incluye ejemplos?
3. Si el test debe contener elementos de distinta dificultad, ¿están los elementos ordenados por dificultad creciente?
4. ¿Se especifican los procedimientos usados en el muestreo de ítems?
5. ¿Tiene el test instrucciones adecuadas para el tratamiento de la adivinación o el azar?
6. Si en un test de personalidad se ha incluido la opción «¿?», ¿se ha especificado el significado de dicha opción?

7. ¿Se ha analizado el sesgo o posible validez diferencial para grupos sociodemográficos?
8. ¿Tiene una acogida favorable por los sujetos a los que va dirigido?
9. ¿No provoca recelos ni suspicacias en los sujetos?
10. ¿Existe en el manual información suficiente sobre las propiedades estadísticas? En el caso de la TRI, ¿están calibrados los ítems?
11. Si el test es de tiempo límite, ¿está especificado?
12. ¿Es sencilla y fiable la forma de corrección y puntuación?
13. ¿Es suficientemente clara la hoja de respuestas, de modo que su cumplimentación no suponga una tarea añadida a la resolución de los problemas del test?
14. ¿Especifica el test las diferentes muestras y el diseño muestral seguido en su tipificación?
15. ¿Refleja el manual la bibliografía utilizada en relación al atributo que mide el test?

2. Teoría Clásica de los Tests (TCT) I: el modelo y la fiabilidad de las puntuaciones

1. El modelo de la TCT

Después de que los ítems son construidos y aplicados a los sujetos, hay que relacionar las respuestas con el constructo que los inspiró. Es de decir, hay que obtener alguna forma de puntuación en el constructo. Esta relación se establece por medio del denominado «modelo de medida». Los modelos de medida de uso extendido en la psicometría responden básicamente a dos formas:

- Los que ponen el acento en la puntuación total obtenida a partir del conjunto de ítems, que se basan en alguna forma de agregación de las puntuaciones de los ítems, normalmente mediante su suma o suma ponderada. Esta aproximación es denominada por Wilson (2005) la *teoría simple de la puntuación.* El modelo que la justifica es la denominada *Teoría Clásica de los Tests* (TCT) y alguna de sus extensiones, como la *Teoría de la Generalizabilidad* (TG).
- Los que ponen el acento en los ítems considerados individualmente, y en los patrones de respuesta. Hay diferentes modelos agrupados de forma general bajo la denominación de *Teoría de la Respuesta al Ítem*.

En el proceso de la construcción del test pueden seguirse ambas aproximaciones, por lo que las dos son tratadas en el texto. En ellas se pretende justificar de algún modo la precisión de las observaciones para evaluar el valor desconocido del constructo latente. Comenzamos por la TCT, la más antigua

y todavía hoy la de uso más generalizado, que se explica en este capítulo y en el capítulo 3; esta teoría es seguida por su extensión más importante, la TG, que se explica en el capítulo 4. En los capítulos 5, 6 y 7, se presentan los modelos y principales conceptos de la TRI.

2. Supuestos y deducciones del modelo

La TCT fue formulada por Spearman (1904a) casi en su totalidad. Su desarrollo se llevó a cabo fundamentalmente durante la primera mitad del siglo pasado y queda reflejada en toda su extensión en el libro de Gulliksen (1950/1987). La teoría fue formalizada en el marco de la estadística del momento, la correlacional Pearsoniana. Lord y Novick (1968) llevaron a cabo una nueva formalización.

El modelo parte de que las respuestas dadas a los ítems del test se combinan en una única puntuación total X_i para cada sujeto *i*. La formalización de la relación de esta puntuación con el constructo se refleja en el siguiente modelo:

$$X_i = V_i + E_i \qquad [2.1]$$

El modelo expresa, simplemente, que la puntuación observada surge de una puntuación verdadera, V_i que es la cantidad que el sujeto posee del atributo más un error de medida, E_i. Como puede observarse, la relación entre X y E es aditiva, dando lugar a un modelo lineal.

Para una persona *s,* la puntuación V_i es constante, es el verdadero estatus del sujeto en el constructo, pero las puntuaciones observadas y los errores son variables aleatorias, ya que diferentes aplicaciones repetidas de un mismo instrumento de medida, aun realizadas bajo las mismas condiciones, proporcionan diferentes puntuaciones observadas. Sucede algo similar a lo que ocurre en la teoría de muestras en estadística, cuando en muestras repetidas de la misma población se obtienen valores diferentes para los estimadores de los parámetros. Puesto que las puntuaciones observadas en un test son las únicas que se pueden obtener empíricamente, las puntuaciones verdaderas y el error deben expresarse por medio de las puntuaciones observadas.

Como todo modelo formal, la TCT necesita de un conjunto de supuestos acerca de la naturaleza de las variables implicadas y de sus relaciones. La TCT se formula a partir de supuestos débiles, que se acomodan a la mayoría de las situaciones prácticas (de ahí que también se la conozca como «modelo débil de la puntuación verdadera»).

Los supuestos del modelo de la TCT se resumen en los siguientes [1]:

$$1)\ E(E_i) = 0 \qquad [2.2]$$

[1] Se establecen los supuestos en términos de covarianzas de las puntuaciones.

El primer supuesto establece que el valor esperado de la variable aleatoria *error de medida* es igual a cero, para una población de personas medidas con el mismo test o para una repetición infinita de medidas realizadas sobre la misma persona.

$$2)\ \sigma_{V_i E_i} = 0 \qquad [2.3]$$

El segundo supuesto establece la falta de relación ente la puntuación verdadera y el error, ello implica que los sujetos con diferentes valores en el atributo se ven afectados por cantidades de error que no tienen que ver con su estatus en el atributo.

$$3)\ \sigma_{E_i E_j} = 0 \qquad [2.4]$$

El tercer supuesto indica que si se aplican dos test distintos, a una población de sujetos, los errores que se cometan al aplicar el primer test no estarán correlacionados con los errores que se cometan al aplicar el segundo test.

$$4)\ \sigma_{V_i E_j} = 0 \qquad [2.5]$$

El cuarto supuesto establece que los errores de medida de un test tampoco están correlacionados con las puntuaciones verdaderas de otro test, ya que en otro caso no se trataría de un error aleatorio, sino de un error relacionado con alguna dimensión medida por segundo test.

Por tanto, el modelo de la TCT considera el *error de medida* como una desviación aleatoria, no sistemática, de la puntuación verdadera. En las aplicaciones de un test también se producen errores sistemáticos, pero no son objeto de estudio de los modelos psicométricos en general, ni de la TCT, en particular.

A partir de los supuestos anteriores, se establecen una serie de deducciones importantes, entre las que cabe destacar las siguientes:

$$E(X_i) = E(V_i) \qquad [2.6]$$

El valor esperado de las puntuaciones observadas es igual al valor esperado de las puntuaciones verdaderas, puesto que el valor esperado de los errores es cero.

$$E(E_i \mid V_i) = 0 \qquad [2.7]$$

La esperanza matemática del error para una subpoblación con la misma puntuación verdadera será cero.

$$\sigma_x^2 = \sigma_v^2 + \sigma_e^2 \qquad [2.8]$$

Puesto que las puntuaciones verdaderas y los errores son independientes, la varianza de las puntuaciones observadas es la suma de la varianza de las puntuaciones verdaderas y de la varianza de los errores de medida. Si no hubiese errores de medida, todas las diferencias entre puntuaciones observadas reflejarían diferencias entre puntuaciones verdaderas y, por tanto, $\sigma_x^2 = \sigma_v^2$.

$$\sigma_{xv} = \sigma_v^2 \qquad [2.9]$$

La covarianza entre las puntuaciones empíricas y verdaderas es igual a la varianza de las puntuaciones verdaderas, puesto que lo que tienen en común las puntuaciones empíricas y las verdaderas es justamente la puntuación verdadera, siendo independientes de los errores.

Una justificación formal de las deducciones anteriores puede encontrarse en Gulliksen (1950/1987) y en Martínez Arias (1995).

Puede definirse ahora la correlación a partir de la covarianza y de las desviaciones típicas de las puntuaciones implicadas, que mediante algunas de las deducciones anteriores, puede establecerse como:

$$\rho_{xv} = \frac{\sigma_v}{\sigma_x} \qquad [2.10]$$

La ecuación anterior que indica la correlación entre *X* y *V* se denomina en la TCT *índice de fiabilidad*.

En estadística es frecuente utilizar la correlación al cuadrado, que es el cociente de la varianza de las puntuaciones verdaderas a la varianza de las puntuaciones observadas.

$$\rho_{xv}^2 = \frac{\sigma_v^2}{\sigma_x^2} \qquad [2.11]$$

Este cociente indica la proporción de varianza de las puntuaciones observadas explicada por su relación lineal con las puntuaciones verdaderas. Las puntuaciones verdaderas explicarían perfectamente las puntuaciones observadas cuando $\rho_{xv}^2 = 1$, es decir, cuando toda la varianza observada se debiera a varianza verdadera. Spearman (1904a) denominó a esta razón *coeficiente de fiabilidad* de las puntuaciones que, aunque es una cantidad al cuadrado, suele denotarse mediante el símbolo ρ_{xx}, por razones que se presentan más adelante.

Una formulación alternativa puede hacerse a partir de la proporción de la varianza de los errores:

$$\rho_{xe}^2 = \frac{\sigma_e^2}{\sigma_x^2} \qquad [2.12]$$

La correlación al cuadrado de las puntuaciones observadas y errores es igual al cociente entre la varianza de ambas variables. Este cociente indica la proporción de varianza de las puntuaciones observadas no explicada por su relación lineal con las puntuaciones verdaderas, sino por el error aleatorio. A partir de esta proporción, puede derivarse otra expresión equivalente para el coeficiente de fiabilidad:

$$\rho_{xv}^2 = 1 - \frac{\sigma_e^2}{\sigma_x^2} = 1 - \rho_{xe}^2 \qquad [2.13]$$

En la expresión anterior puede verse el efecto de los errores sobre el coeficiente de fiabilidad. Por tanto, cuando la varianza de los errores es pequeña con relación a la varianza observada, el valor del coeficiente de fiabilidad será elevado.

3. Concepto de fiabilidad

Cuando se utiliza cualquier instrumento para medir, ya sean variables físicas o psicológicas, la medida resultante siempre está afectada por un error aleatorio. Si aplicamos el mismo instrumento de medida a los mismos sujetos en varias ocasiones, el resultado no será siempre el mismo para cada uno de ellos, sino que proporcionará diferentes estimaciones de una ocasión a otra. Este hecho revela una falta de fiabilidad en el instrumento. Afortunadamente, las mediciones realizadas en diferentes aplicaciones también muestran consistencias. Cuando aplicamos un test de motivación a un conjunto de sujetos, un sujeto que obtenga la puntuación más alta en la primera medición, en una segunda medición estará entre los que obtenga una puntuación más alta, aunque no sea exactamente la mayor. Esta tendencia a la *consistencia* de un conjunto de medidas es lo que se denomina *fiabilidad.*

En la medición de variables psicológicas el problema de la consistencia es más complejo que en la medición de variables físicas, ya que en repeticiones del mismo test, el resultado puede variar tanto por falta de fiabilidad o precisión del instrumento como por otros aspectos del propio sujeto o del contexto en el que se realiza la medición. En estos dos últimos casos la falta de estabilidad de las puntuaciones no puede atribuirse a la falta de fiabilidad del instrumento. Existen, por tanto, otro tipo de errores, que pueden explicarse por la teoría psicológica, denominados *errores sistemáticos*, que no son objeto de estudio de la teoría de la fiabilidad. Hay variaciones que son explicables por la teoría del constructo que estamos midiendo (por ejemplo, si entre la primera aplicación de un test y la segunda ha tenido lugar un proceso de aprendizaje o de maduración) y otras por cambios del contexto de la medida (por ejemplo, si se aplica el instrumento de medida al principio de una jornada, donde el sujeto está menos cansado,

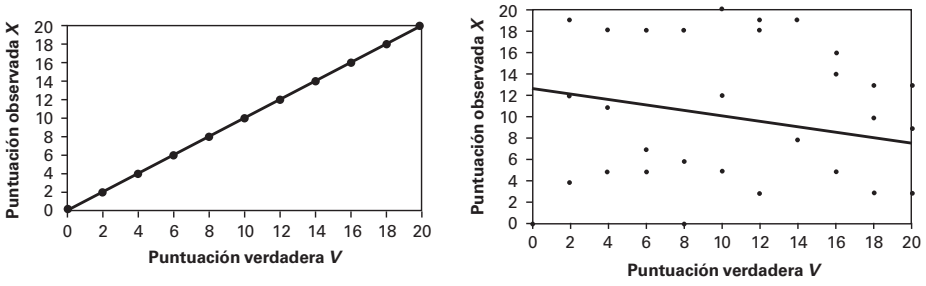
o al final, donde pueden aparecer efectos de fatiga). En todos los casos las variaciones son sistemáticas y predecibles. Por el contrario, la fiabilidad de la medida se refiere a variaciones en las puntuaciones que se producen sin razones teóricas y/o empíricas que justifiquen el cambio, por lo que las diferencias encontradas se deben a errores aleatorios asociados al proceso de medición.

El nivel de fiabilidad de las puntuaciones es fundamental, tanto para el uso de los tests con fines de investigación, como para su utilización en cualquier proceso de decisión. En el caso de la investigación, no podemos hacer predicciones ni inferencias a partir de puntuaciones con baja fiabilidad, ya que la fiabilidad limita el valor de la correlación que podemos obtener entre dos medidas y afecta a los diferentes estadísticos calculados a partir de los datos, como se verá a lo largo del capítulo. En la utilización de los test para la toma de decisiones, a menor fiabilidad mayor probabilidad de tomar una decisión incorrecta sobre el sujeto.

La razón de definir el coeficiente de fiabilidad como lo hizo Spearman (véase en la ecuación 2.12) desde la teoría de la regresión lineal fue porque representa una buena forma de establecer el grado de aproximación entre las puntuaciones verdaderas y las observadas. Si recordamos la teoría de la regresión de estadística, para cada puntuación verdadera puede obtenerse una distribución de puntuaciones observadas X, de tal modo que $E(X_i \mid V_i) = V_i$. Bajo este modelo hablamos del coeficiente de determinación, ρ^2_{xv} o proporción de la variación de X explicada por V, es decir, por las puntuaciones verdaderas, definido como $\rho^2_{xv} = \sigma^2_v/\sigma^2_x$, que es la misma expresión dada en la ecuación 2.12 o en términos de los errores como en la ecuación 2.13.

El coeficiente así definido se denomina coeficiente de fiabilidad del test y puede considerarse una buena definición de la fiabilidad del instrumento, como se observa claramente en los gráficos representados en la figura 2.1.

Figura 2.1. Representación de la relación lineal entre puntuaciones verdaderas y empíricas y el diagrama de dispersión para dos casos extremos

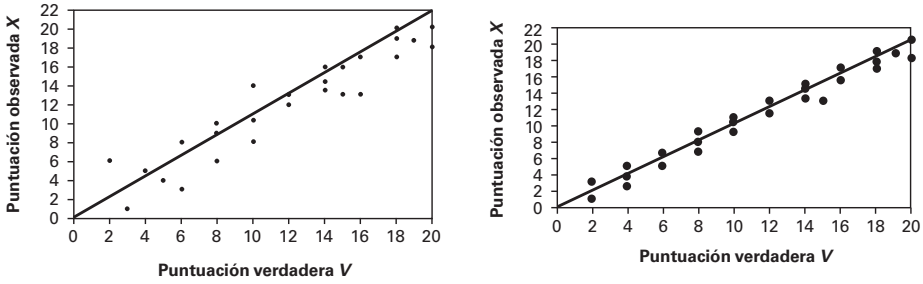


(a) (b)

En la figura 2.1 (a) se representa el hipotético caso en el que una puntuación verdadera V_i siempre da lugar a la misma puntuación observada, X_i, y esto es así para las distintas puntuaciones verdaderas. En este caso, todas las puntuaciones se encuentran en la recta de regresión y en términos del modelo significa que $E = X - V = 0$, no existiendo discrepancias entre las puntuaciones observadas y las verdaderas, ni entre dos mediciones diferentes de la misma puntuación verdadera. En estos casos todos los errores son 0, igual que su varianza y la situación reflejaría una fiabilidad perfecta. En la figura 2.1 (b) nos encontramos el caso opuesto, ya que las diferentes puntuaciones verdaderas dan lugar a distribuciones similares en rango de puntuaciones observadas, es decir, que sujetos con puntuaciones verdaderas diferentes pueden obtener la misma puntuación empírica y sujetos con la misma puntuación verdadera pueden obtener puntuaciones empíricas muy diferentes. Como para la misma puntuación verdadera cualquier diferencia $X - V$ se debe a errores de medida, podemos decir que $\sigma_x^2 = \sigma_e^2$ para todo el rango de puntuaciones verdaderas y, en este caso, toda la varianza observada se debe a varianza de los errores. En estas condiciones la fiabilidad del instrumento de medida será 0.

En la mayor parte de los casos, la situación será similar a la encontrada en la figura 2.2. En (a) puede observarse que los sujetos con puntuaciones verdaderas bajas tienden a tener puntuaciones observadas bajas, con puntuaciones medias, puntuaciones observadas medias, y con puntuaciones verdaderas elevadas, puntuaciones observadas también elevadas. No obstante, existen algunas discrepancias entre puntuaciones observadas para una misma puntuación verdadera, debida a errores de medida. La dispersión de las puntuaciones en torno a la puntuación verdadera se debe a errores y su varianza impide que la correlación sea perfecta. En (b) la dispersión es menor, lo que sugiere que los errores o discrepancias con las puntuaciones verdaderas son menores y también su varianza.

Figura 2.2. Representación de la relación lineal entre puntuaciones verdaderas y empíricas y el diagrama de dispersión para dos casos habituales



(a) (b)

4. Relaciones del modelo basadas en el supuesto de paralelismo

La imposibilidad de calcular empíricamente el coeficiente de fiabilidad llevó a Spearman al concepto de *tests o medidas paralelas*. Dos tests, cuyas puntuaciones observadas se denotan como X y X', son paralelos en la concepción de Spearman, si se dan en ellos los siguientes supuestos:

- Igualdad de las puntuaciones verdaderas, es decir:

$$X = V + E$$
$$X' = V + E'$$

- Igualdad de las varianzas de los errores, $\sigma_e^2 = \sigma_{e'}^2$.
- Los dos supuestos anteriores conducen a la igualdad de medias y de varianzas de las puntuaciones observadas

$$\mu_x = \mu_{x'} \text{ y } \sigma_x^2 = \sigma_{x'}^2$$

Fácilmente se demuestra que la correlación entre dos tests paralelos es igual a la razón de la varianza verdadera a la varianza de las puntuaciones observadas, es decir:

$$\rho_{xx'} = \frac{\sigma_v^2}{\sigma_x^2} = \rho_{vx}^2 \qquad [2.14]$$

Por tanto, los supuestos del paralelismo permiten llegar a una estimación empírica del coeficiente de fiabilidad, ya que ahora las puntuaciones de las dos formas X y X' pueden obtenerse empíricamente y calcular con ellas el coeficiente de fiabilidad. Esta es la razón de denotar este coeficiente como $\rho_{xx'}$ o a veces como ρ_{xx}.

Una vez calculado el coeficiente de fiabilidad, pueden estimarse los valores de las varianzas verdadera y error:

$$\sigma_v^2 = \sigma_x^2 \rho_{xx'} \qquad [2.15]$$

$$\sigma_e^2 = \sigma_x^2 (1 - \rho_{xx'}) \qquad [2.16]$$

La desviación típica de los errores, obtenida como la raíz cuadrada de la expresión anterior, se denomina error típico de medida.

$$\sigma_e = \sqrt{\sigma_x^2 (1 - \rho_{xx'})} = \sigma_x \sqrt{(1 - \rho_{xx'})} \qquad [2.17]$$

Dentro del modelo lineal en el que se inscribe la teoría clásica de los tests, la varianza de los errores se supone que es homocedástica, es decir, que es *la*

misma en todos los niveles de las puntuaciones. Lo mismo es aplicable al error típico de medida.

Otras dos relaciones interesantes, cuando se dispone de conjuntos de varios tests paralelos, $X_1, X_2, \ldots, X_j, \ldots, X_p$, son las siguientes:

- Las correlaciones entre cualquier par de tests de un conjunto de tests paralelos son iguales, e iguales al coeficiente de fiabilidad: $\rho_{12} = \ldots = \rho_{1p} = \ldots = \rho_{p,p-1}$.
- Las correlaciones de cualquiera de los tests del conjunto con una variable externa al mismo, Y, también son iguales entre sí: $\rho_{1y} = \ldots = \rho_{p-1,y} = \ldots \rho_{py}$.

Puesto que el coeficiente de fiabilidad se trata de un coeficiente de correlación al cuadrado de X con V o de la correlación empírica entre dos tests paralelos, sus valores deben oscilar entre 0 y 1, alcanzando el valor de 1 solamente cuando no hay errores de medida y el de 0, cuando toda la varianza observada se debe a errores de medida.

La correlación entre dos formas paralelas es, evidentemente, una forma de estimar la fiabilidad del test, pero como, en la práctica, el supuesto de paralelismo es muy difícil de conseguir, se han propuesto otros conceptos de equivalencia entre medidas que suponen una relajación del supuesto de paralelismo: medidas equivalentes en puntuación verdadera (o tau-equivalentes), medidas esencialmente equivalentes en puntuación verdadera (o esencialmente tau-equivalentes) y medidas linealmente equivalentes en puntuación verdadera (o congenéricas).

Las medidas *equivalentes en puntuación verdadera*, o *tau-equivalentes*[2], suponen la misma puntuación verdadera pero admite diversidad en la varianza error.

$$X_j = V + E_j$$
$$X_h = V + E_h$$

Por tanto, también se cumple la igualdad de medias y de varianzas de las puntuaciones verdaderas. Pero no se cumple la igualdad de las varianzas de los errores ni de las puntuaciones empíricas.

Las medidas *esencialmente equivalentes en puntuación verdadera*, o esencialmente tau-equivalentes, permiten que las puntuaciones verdaderas de ambas medidas difieran en una constante aditiva.

$$X_j = V_j + E_j$$
$$X_h = V_h + E_h$$

donde: $V_h = V_j + a_{jh}$.

[2] De τ, letra griega correspondiente a la palabra inglesa verdadera: *true*.

Las medidas *linealmente equivalentes en puntuación verdadera*, o congenéricas, permiten que la puntuación verdadera de una de ellas sea transformación lineal de la otra.

$$X_j = V_j + E_j$$
$$X_h = V_h + E_h$$

donde: $V_h = b_{jh}V_j + a_{jh}$.

Por tanto, no se cumplen ni la igualdad de varianza ni la igualdad de medias de ambas medidas, aunque se puede obtener la transformación de una en otra.

Para estos tres tipos de equivalencia, en los que no se cumple el supuesto de paralelismo, las puntuaciones empíricas son distintas, por tanto:

- Las correlaciones entre cualquier par de tests de un conjunto de test equivalentes serán diferentes y, por tanto, obtendremos diferentes estimaciones del coeficiente de fiabilidad.
- Las correlaciones de cualquiera de los tests de un conjunto de tests equivalentes con una variable externa también pueden ser diferentes.

Una explicación más extensa sobre estos tipos de modelos puede encontrarse en Traub (1994).

5. Factores que afectan al coeficiente de fiabilidad

La estimación del coeficiente de fiabilidad, calculada a partir de datos empíricos, está afectada por varios factores cuyo efecto analizaremos a lo largo de este apartado: 1) variabilidad del grupo y 2) longitud del test.

5.1. Fiabilidad y variabilidad de la muestra

Como el lector recordará de estadística, el coeficiente de correlación es sensible al rango o variabilidad de las variables, en el sentido de que, manteniendo constantes el resto de los factores, a menor variabilidad, menor correlación. Esto mismo sucede con el coeficiente de fiabilidad. En circunstancias particulares, como cuando se calcula con muestras de variabilidad menor que la que tiene la variable en la población, el problema suele denominarse de *restricción del rango*. El supuesto de la homocedasticidad, o igualdad de la varianza de las puntuaciones a lo largo de todo el rango de la variable, ha permitido a la TCT obtener unas ecuaciones para estimar el valor del coeficiente de fiabilidad en muestras distintas a aquellas en las que se ha calculado.

Sean dos muestras que denotamos como *j* y *k*, respectivamente, con varianzas σ_j^2 y σ_k^2 y coeficientes de fiabilidad $\rho_{jj'}$ y $\rho_{kk'}$. Suponiendo que el coeficiente se ha calculado en la muestra *j* y si conocemos la varianza de la muestra o población *k*, puede obtenerse para ésta una estimación de su coeficiente de fiabilidad:

$$\rho_{kk'} = 1 - \frac{\sigma_j^2}{\sigma_k^2}(1 - \rho_{jj'}) \qquad [2.18]$$

Lord y Novick (1968) advierten que esta estimación debe usarse con cautela, dado que el supuesto de homocedasticidad de los errores no siempre es sostenible en la práctica.

El problema de la restricción del rango también afecta a las correlaciones con variables externas. Gulliksen (1950/1987) presenta ecuaciones para este problema que se expondrán en el capítulo sobre validez.

5.2. Fiabilidad y longitud: fórmula de la profecía de Spearman-Brown

Un aspecto muy influyente en el valor del coeficiente de fiabilidad es la *longitud* del test o número de ítems que lo componen. Spearman y Brown, de forma independiente, propusieron en 1910 una fórmula, conocida a veces como la *fórmula de la profecía,* que permite estimar cambios en el coeficiente de fiabilidad producidos por la adición de ítems paralelos a los existentes. Este procedimiento también puede aplicarse si queremos conocer en cuánto se reducirá la fiabilidad eliminando ítems del test. Una aplicación frecuente es su utilización como corrección en el procedimiento de cálculo de la fiabilidad a partir en la división de un test en dos mitades, que consideraremos en el siguiente apartado. La deducción de las ecuaciones puede encontrarse en Martínez Arias (1995).

Sea un test con un coeficiente de fiabilidad $\rho_{xx'}$ y sea *k* un valor multiplicador de su longitud (p. ej.: si duplicamos, $k = 2$; si triplicamos, $k = 3$; si pasamos de una longitud de 10 ítems a 15, $k = 1{,}5$ y si reducimos de 45 a 15, $k = 1/3$). Denotamos por ρ_k el coeficiente de fiabilidad del test obtenido multiplicando por *k* la longitud del original.

Los cambios en el coeficiente de fiabilidad se producen por cambios en las varianzas verdadera, error y observada de las puntuaciones. Veamos cuáles son los efectos de multiplicar por *k* la longitud, añadiendo elementos paralelos:

1. Sea $\overline{X}$ la media de las puntuaciones del test original formado por *n* elementos. Si se multiplica la longitud por *k* añadiendo elementos paralelos, sería lo mismo que si se sumasen las puntuaciones de *k* tests paralelos, por lo que la media del test modificado sería $k\overline{X}$.

2. Efectos sobre las varianzas observadas, verdaderas y error. Puede demostrarse (véase Martínez Arias, 1995, p. 102) que el efecto de multiplicar la longitud por k produce los cambios siguientes en estas varianzas, donde σ_x^2, σ_v^2 y σ_e^2 son las varianzas observada, verdadera y error del test original de n elementos y σ_{kx}^2, σ_{kv}^2 y σ_{ke}^2 las correspondientes varianzas empírica, verdadera y error del test multiplicado por k en longitud, añadiendo elementos paralelos:

$$\sigma_{kx}^2 = k\sigma_x^2 + k(k-1)\rho_{xx'}\sigma_x^2$$
$$\sigma_{kv}^2 = k^2\sigma_v^2$$
$$\sigma_{ke}^2 = k\sigma_e^2$$

Sustituyendo los valores de σ_k^2 y σ_{kv}^2 en la definición de fiabilidad como razón de ambas varianzas, y simplificando, se llega a la siguiente ecuación para estimar el coeficiente de fiabilidad en longitud multiplicada por k:

$$\rho_k = \frac{k\rho_{xx'}}{[1 + (k-1)\rho_{xx'}]} \qquad [2.19]$$

El coeficiente de fiabilidad aumenta siempre que se añadan ítems paralelos, teniendo como límite 1, en el caso de que se añadiesen infinitos ítems.

En la figura 2.3, adaptada de Gulliksen (1950), pueden observarse los cambios en la fiabilidad con cambios en la longitud.

De esta ecuación general puede deducirse otra que permite estimar el número de elementos requerido para alcanzar una cierta fiabilidad

$$k = \frac{\rho_k(1 - \rho_{xx'})}{\rho_{xx'}(1 - \rho_k)} \qquad [2.20]$$

donde los términos tienen el mismo significado anterior. K nos da el coeficiente por el que habrá de multiplicar la longitud actual del test para llegar a una nueva longitud n' en la que el coeficiente de fiabilidad alcanza el valor deseado, es decir, $n' = kn$ y el número de elementos a añadir será $n' - n$, redondeado al entero más próximo.

El caso de duplicar la longitud ($k = 2$) es un caso particular de la situación general reflejada en la ecuación de Spearman-Brown. Los cambios en las varianzas observada, verdadera y error serán ahora:

$$\sigma_{xk=2}^2 = 2\sigma_x^2(1 + \rho_{xx'})$$