Introducción a la Psicometría

Teoría clásica y TRI

José Muñiz

PIRÁMIDE

Introducción a la Psicometría

Teoría clásica y TRI

JOSÉ MUÑIZ
CATEDRÁTICO DE PSICOMETRÍA DE LA UNIVERSIDAD DE OVIEDO

Introducción a la Psicometría

Teoría clásica y TRI

EDICIONES PIRÁMIDE

COLECCIÓN «PSICOLOGÍA»

Director:
Francisco J. Labrador
Catedrático de Modificación de Conducta
de la Universidad Complutense de Madrid

Edición en versión digital

Está prohibida la reproducción total o parcial de este libro electrónico, su transmisión, su descarga, su descompilación, su tratamiento informático, su almacenamiento o introducción en cualquier sistema de repositorio y recuperación, en cualquier forma o por cualquier medio, ya sea electrónico, mecánico, conocido o por inventar, sin el permiso expreso escrito de los titulares del copyright.

© José Muñiz, 2018
© Primera edición electrónica publicada por Ediciones Pirámide (Grupo Anaya, S. A.), 2018
Para cualquier información pueden dirigirse a piramide_legal@anaya.es
Juan Ignacio Luca de Tena, 15. 28027 Madrid
Teléfono: 91 393 89 89
www.edicionespiramide.es
ISBN digital: 978-84-368-3933-3

A mis hijas Laura y Marta.
A mi mujer Alejandra.

Índice

Prólogo 13

1. Introducción 15

1. Orígenes y desarrollo de la teoría de los test 16
2. Modelo lineal clásico 21
3. Deducciones inmediatas del modelo 22

2. Fiabilidad 25

1. Coeficiente de fiabilidad 26
2. Estimación empírica del coeficiente de fiabilidad 26
3. Estimación de las puntuaciones verdaderas 28
4. Fiabilidad de las diferencias 31
5. Tipos de errores de medida 32
6. Factores que afectan a la fiabilidad 33
   - 6.1. Fiabilidad y variabilidad 33
   - 6.2. Fiabilidad y longitud 34
   - 6.3. Fiabilidad y nivel de las puntuaciones en el test 36
7. Coeficiente *alfa* (α) 39
   - 7.1. Concepto y fórmula 39
   - 7.2. Casos particulares de α 40
   - 7.3. Cálculo de α mediante análisis de varianza 43
   - 7.4. Coeficiente *beta* (β) 44
   - 7.5. Coeficientes basados en el análisis factorial de los ítems 45
   - 7.6. Inferencias sobre α 46
8. Teoría de la generalizabilidad 51
   - 8.1. Fuentes de error 51
   - 8.2. Conceptos básicos 53
   - 8.3. Diseños de recogida de datos 56
   - 8.4. Coeficiente de generalizabilidad 58
   - 8.5. Estudios de generalizabilidad y estudios de decisión 60
   - 8.6. Error típico de medida 61
   - 8.7. Diseños de dos facetas 62
9. Fiabilidad de los test referidos al criterio 63
   - 9.1. Definición 63
   - 9.2. Métodos de estimación de la fiabilidad 65

© Ediciones Pirámide

9.3. Establecimiento del punto de corte 75
9.4. Fiabilidad interjueces 84
9.5. Comentarios finales 86
Ejercicios 87

3. Validez 101

1. Concepto 101
1.1. Evidencias de validez 102
2. Validez y fiabilidad 109
2.1. Fórmulas de atenuación 109
2.2. Valor máximo del coeficiente de validez 111
2.3. Validez y longitud del test 112
3. Validez y variabilidad 113
3.1. Dos variables 113
3.2. Tres variables 115
3.3. Caso general 116
4. Validez y predicción 117
4.1. Regresión simple 117
4.2. Regresión múltiple 120
4.2.1. Modelo 121
4.2.2. Correlación múltiple 125
4.2.3. Correlación parcial y semiparcial 128
4.2.4. Variables moduladoras y supresoras 130
4.3. Validez y decisión 130
4.3.1. Índices de validez 131
4.3.2. Incidencia del punto de corte en los tipos de errores 133
4.3.3. Curvas ROC 134
4.4. Selección y clasificación 135
4.4.1. Modelos de selección 135
4.4.2. Utilidad de la selección 137
4.4.3. Clasificación 139
Ejercicios 141

4. Análisis de los ítems 149

1. Índice de dificultad 149
2. Índice de discriminación 150
2.1. Cálculo 151
2.2. Relación con algunos parámetros del test 153
3. Índice de validez 155
3.1. Relación con los parámetros del test 155
4. Análisis de las alternativas incorrectas 157
4.1. Número óptimo de alternativas 157
5. Corrección del azar 158
6. Calificación del conocimiento parcial 160
7. Funcionamiento diferencial de los ítems 161
7.1. Introducción 161
7.2. Concepto 161
7.3. Métodos de evaluación 163
7.3.1. χ^2 de los aciertos 163
7.3.2. χ^2 global 165
7.3.3. Método *delta* 166

© Ediciones Pirámide

7.3.4. Mantel-Haenszel 168
7.3.5. Índice de estandarización 172
7.3.6. Comentarios finales 173
Ejercicios 174

5. Transformación de las puntuaciones 177

1. Percentiles 177
2. Puntuaciones típicas 178
2.1. Típicas derivadas 178
2.2. Típicas normalizadas 179
3. Edad 180
Ejercicios 181

6. Equiparación de las puntuaciones 183

1. Diseños 184
2. Métodos 184

7. Teoría de respuesta a los ítems 187

1. Objetivos 187
2. Supuestos 188
2.1. Curva característica de los ítems 188
2.2. Unidimensionalidad e independencia local 192
Ejercicios 193
3. Modelos 195
3.1. Modelo logístico de un parámetro (modelo de Rasch) 195
3.2. Modelo logístico de dos parámetros 197
3.3. Modelo logístico de tres parámetros 198
3.4. Modelos de ojiva normal 200
3.5. Orígenes y desarrollo de la TRI 201
Ejercicios 206
4. Aplicación de los modelos 207
4.1. Comprobación de la unidimensionalidad 207
4.2. Elección del modelo 209
4.3. Estimación de los parámetros 210
4.4. Ajuste del modelo 214
4.5. Invarianza de los parámetros 218
Ejercicios 220
5. Métrica de theta (θ) 222
5.1. Transformaciones admisibles de θ 223
5.2. Transformaciones de P(θ): *logits* 224
5.3. Otras transformaciones 226
Ejercicios 226
6. Curva característica del test 227
6.1. Definición 227
6.2. Puntuaciones verdaderas en el test 228
6.3. Curva característica de la persona 230
Ejercicios 231
7. Función de información 233
7.1. Error típico de estimación de θ 233

© Ediciones Pirámide

7.2. Función de información del test 233
7.3. Funciones de información de los modelos logísticos 235
7.4. Función de información de los ítems 235
7.5. Información máxima 236
7.6. Ponderación óptima de los ítems 237
7.7. Eficiencia relativa de dos test 237
7.8. Función de información y transformaciones de θ 239
Ejercicios 239
8. Bancos de ítems 241
8.1. Concepto y desarrollo 241
9. Equiparación de puntuaciones 243
9.1. Concepto y técnicas 243
Ejercicios 247
10. Funcionamiento diferencial de los ítems 248
10.1. Concepto 248
10.2. Evaluación 250
Ejercicios 260
11. Test adaptativos informatizados 261
11.1. Concepto 261
11.2. Desarrollo 261
Ejercicios 264

8. Fases para la construcción de un test 265

1. Marco general 265
2. Definición de la variable medida 267
3. Especificaciones 268
4. Construcción de los ítems 268
4.1. Tipos de ítems 269
4.2. Directrices para la construcción de ítems de elección múltiple 274
5. Edición 279
6. Estudios piloto 279
7. Selección de otros instrumentos de medida 280
8. Aplicación del test 280
9. Propiedades psicométricas 281
10. Versión final del test 282

9. Utilización de los test 283

1. Estrategias para mejorar el uso de los test 283
2. Formación de los usuarios 284
3. Estándares técnicos 285
4. Preparación para los test 286
5. Utilización de los datos de los test 287
6. Modelo de evaluación de la calidad de los test 288

10. Mirando hacia el futuro 311

Apéndice 315

Tablas estadísticas 329

Referencias bibliográficas 345

© Ediciones Pirámide

Prólogo

Los test son instrumentos de medida que utilizan habitualmente los psicólogos para obtener información que les ayude a tomar decisiones bien fundamentadas. Estas decisiones pueden tener repercusiones importantes en la vida de las personas; por tanto, es clave que los test utilizados cumplan unos requisitos de calidad psicométrica demostrables. Precisamente de eso trata el libro que tiene entre las manos, de mostrar cuáles deben ser las propiedades psicométricas de los test para que puedan ser utilizados con garantía por los psicólogos. Queridos lectores, lo que van a encontrar en las páginas que siguen es una introducción al campo de la psicometría, como bien anuncia, sin sorpresas, el título del libro. Está pensando para alguien que aún no sabe, pero que pretende saber; por eso la filosofía del libro es iniciar a los estudiantes, profesionales y en general personas interesadas en la evaluación psicométrica. Los contenidos se ajustan a la materia que los estudiantes de grado de psicología, educación y ciencias sociales y de la salud deben aprender en un semestre dedicado a la psicometría. También los ya iniciados pueden repasar y actualizar sus conocimientos, pues nunca viene mal volver a hollar los caminos andados y descubrir detalles que se nos habían pasado desapercibidos. A todos vosotros va dirigido este libro introductorio, que pretende ser de amigable lectura, pues no supone grandes conocimientos estadísticos o psicométricos previos. Si lo leyereis con provecho, os abrirá las puertas a otros textos más avanzados, de los cuales hay actualmente abundancia tanto en nuestra propia lengua española como en inglés, y que se irán citando en su momento para animar a profundizar en los temas tratados. El grueso del libro procede de la fusión y actualización de otros dos previos del autor, publicados en esta misma editorial Pirámide, uno sobre teoría clásica y otro sobre teoría de respuesta a los ítems. Esta fusión ha permitido reunir lo esencial de ambos en un solo volumen, evitando redundancias y actualizando lo hecho, aparte de incluir nuevos temas y suprimir otros. Del enfoque clásico se expone lo fundamental: el modelo, la fiabilidad, la validez, el análisis de los ítems, la transformación de las puntuaciones y la equiparación; sin olvidarnos de los test referidos al criterio y la teoría de la generalizabilidad. En cuanto a la teoría de respuesta a los ítems, tras exponer su lógica y los avances que supone sobre el enfoque clásico, se presentan los principales modelos, la curva característica del test, las funciones de información de los ítems y del test, los bancos de ítems, el funcionamiento diferencial y los test adaptativos informatizados. Finaliza el libro con tres apartados de gran interés aplicado, como son las fases o etapas para construir un test, los problemas implicados en la utilización práctica de los test y unas reflexiones finales sobre el futuro de la evaluación psicométrica. La idea que inspira todo el libro es dar una visión comprensiva, no especializada pero rigurosa, sobre el estado actual de la evaluación psicométrica, sin la cual no es posible un ejercicio profesional basado en evidencias, fieles a la idea de que una evaluación rigurosa es la base de un diagnóstico preciso que a su vez permita una intervención eficaz. Si falla la evaluación, todo lo demás se viene abajo.

El libro no hubiese sido posible sin la ayuda de tantas y tantas personas, unas de forma explícita y otras implícita, que tanto monta. En primer lugar, la fuerza motriz del libro son los alumnos de muchas

© Ediciones Pirámide

partes del mundo a los que vengo impartiendo psicometría desde hace ahora cuarenta años, con especial mención para los de las tres universidades españolas en las que tuve el honor de trabajar: Complutense de Madrid, Baleares y Oviedo. Son ellos quienes con sus dudas y preguntas incisivas me ayudan y estimulan a tratar de presentar la materia de forma amena y comprensible. No sé si lo habré logrado. Muchas gracias también a los compañeros de nuestro grupo de investigación de psicometría: Marcelino Cuesta, Yolanda de la Roca, Rubén Fernández, Eduardo Fonseca, Eduardo García-Cueto, Elena Govorova, Luis Manuel Lozano, Teresa Martínez, Víctor Martínez-Loredo, Fernando Menéndez, Ignacio Pedrosa, Elsa Peña, Francisco Prieto, Javier Suárez y Pamela Woitschach. Nuestras sesiones y discusiones para sacar adelante los proyectos constituyen para mí una verdadera fuente de aprendizaje y motivación. He tenido muchos maestros, caminamos a hombros de gigantes, pero no quiero dejar de citar a dos de quienes más he aprendido directamente en muchos sentidos: Mariano Yela y Ronald Hambleton. Sus profundos conocimientos psicométricos, unidos a su talante personal y bonhomía, fueron una bendición para mí; muchas gracias, maestros. Nada saldría adelante sin el apoyo decidido de la directora editorial Inmaculada Jorge, que tuvo claro el proyecto desde el principio y me animó a ello. Se agradece sinceramente. Por supuesto, la familia siempre está al quite: a mis hijas Laura y Marta y a mi mujer Alejandra va dedicado el libro. Perdonad todos los demás no citados, sé que estáis ahí, como bien nos enseñó Bertolt Brecht en su bello poema «Preguntas de un obrero ante un libro». Muchas gracias a todos por todo. Espero que el libro resulte de interés y ayude a penetrar en los arcanos de la psicometría, pues de eso se trata.

© Ediciones Pirámide

1 Introducción

La psicometría puede definirse en términos generales como el conjunto de métodos, técnicas y teorías implicadas en la medición de las variables psicológicas. Como su nombre indica, trataría de todo aquello relacionado con la medición de lo psicológico. Ahora bien, de la medición de lo psicológico se ocupa también cualquier otro acercamiento riguroso al estudio del comportamiento humano; lo que constituiría lo específico de la psicometría sería su énfasis y especialización en aquellas propiedades métricas exigibles a las mediciones psicológicas independientemente del campo sustantivo de aplicación y de los instrumentos utilizados. Así, por ejemplo, aspectos como la fiabilidad o la validez de las mediciones, por citar dos de los más conocidos, constituyen requisitos exigibles para cualquier evaluación psicológica, sea cual fuere su ámbito de aplicación y enfoque. Este tipo de especialización de la psicometría en las categorías métricas que atraviesan los distintos campos sustantivos de la psicología da lugar a que los tipos de contenidos sobre los que trabajan los psicómetras resulten bastante amplios y variados.

Una ojeada a los congresos organizados por las sociedades psicométricas europea y americana, así como a las revistas científicas del área, permite articular la mayor parte de la temática psicométrica en torno a cinco grandes bloques: teoría de la medición, que abarcaría todo lo relativo a la fundamentación teórica de la medida; teoría de los test, donde se explicitan la lógica y los modelos matemáticos subyacentes a la construcción y uso de los test; escalamiento psicológico, que aborda la problemática inherente al escalamiento de estímulos psicológicos; escalamiento psicofisico, que hace lo propio con los estímulos fisicos, y técnicas multivariadas, que junto con el resto de tecnología estadística resultan imprescindibles para la construcción y análisis de los instrumentos de medida.

Como se puede observar, el campo de referencia del término «psicometría» es amplísimo, y no es nuestro objetivo entrar aquí a realizar una descripción o definición precisa de los bloques citados, cada uno de los cuales está a su vez altamente especializado y estructurado en subáreas. En este libro nos centraremos exclusivamente en la teoría de los test, abordando los dos grandes enfoques, el clásico y la teoría de respuesta a los ítems (TRI). Solo subrayar que el término «psicometría» es mucho más genérico y amplio que el de teoría de los test, con el que erróneamente se le identifica a menudo, tomando la parte por el todo. A partir de los años sesenta se populariza también el término afín de «psicología matemática», utilizado para denominar aquellos trabajos caracterizados por un acercamiento formalizado a los problemas psicológicos, lo cual es básicamente coincidente con lo que se entendía por psicometría. De hecho, Thurstone (1937) utiliza el término «psicología matemática» para caracterizar en pocas palabras el objeto de la sociedad psicométrica americana por él fundada. Lo más específico y diferencial de la psicología matemática respecto de la psicometría serán los modelos matemáticos elaborados para áreas específicas de la psicología, tales como aprendizaje, memoria, percepción, lenguaje, pensamiento, interacción social, etc., que proliferan a partir de los años cincuenta y sesenta, y que son en realidad la razón fundamental que da sentido y apoya la nueva denominación de psicología matemática diferenciada de

© Ediciones Pirámide

la anterior de psicometría. Como ocurriera antes en psicometría, se publican textos con esa denominación genérica (Atkinson, 1964; Coombs, Dawes y Tversky, 1970; Krantz, Atkinson, Luce y Suppes, 1974; Laming, 1973; Luce, Bush y Galanter, 1963; Restle y Greeno, 1970), que luego desaparecerá para dar paso a las subáreas especializadas, aparecen sociedades científicas de psicólogos matemáticos con reuniones y congresos propios desde 1967 y se publican revistas, con mención especial para el *Journal of Mathematical Psychology* y *British Journal of Mathematical and Statistical Psychology.* Que ambos términos, «psicometría» y «psicología matemática», denotan campos muy solapados se comprueba empíricamente al observar que los especialistas publican sus trabajos en las revistas de uno y otro campo y asisten indistintamente a los congresos y reuniones de las respectivas sociedades científicas. Considerar, como hacen Estes (1975) o Greeno (1980), la psicometría como la parte de la psicología matemática dedicada a todo lo relacionado con la medida es tan razonable como considerar la psicología matemática aquella rama de la psicometría dedicada a los modelos matemáticos de procesos psicológicos, pero, eso sí, aquella resulta menos respetuosa con la historia de la psicología, ya que fue «psicometría» el término que originariamente se utilizó para tales menesteres. Así, Thurstone (1937), en su conferencia para la primera reunión anual de la sociedad psicométrica que tuvo lugar en 1936, señala como objetivo de la sociedad «estimular el desarrollo de la psicología como ciencia cuantitativa y racional. O lo que más brevemente puede denominarse psicología matemática». Añade, además, algo que se olvida demasiado a menudo: «A la larga seremos juzgados por la significación, relevancia y consistencia de los principios psicológicos que descubramos», dejando claro desde el principio que la formalización y matematización propias del enfoque psicométrico están al servicio de los problemas psicológicos que tratan de resolver, pero no constituyen un fin en sí mismo para la psicometría.

Quede ahí este breve apunte terminológico para evitar confusiones y desasosiegos al lector que los topare en su deambular por los reales de la psicología, aunque no lo hará con mucha frecuencia, pues, como ya se ha señalado, debido a su amplitud, se han vuelto ambos demasiado genéricos y lo habitual será encontrarse con las denominaciones específicas de las distintas subáreas. Como modesto ejemplo, en este libro se tratará de las teorías de los test, tanto el enfoque clásico como la teoría de respuesta a los ítems.

1. ORÍGENES Y DESARROLLO DE LA TEORÍA DE LOS TEST

El nacimiento formal de la teoría de los test puede ubicarse en los primeros trabajos de Spearman (1904, 1907, 1913), en los que establece los fundamentos de la teoría clásica de los test (TCT). El objetivo central era encontrar un modelo estadístico que fundamentase adecuadamente las puntuaciones de los test y permitiera la estimación de los errores de medida asociados a todo proceso de medición. El modelo lineal clásico propuesto por Spearman, que luego se abordará con cierto detalle, destaca por su sencillez matemática y enjundia psicológica, lo que le garantizará una larga vida. Asume que la puntuación empírica de una persona en un test (X) consta de dos componentes aditivos: uno, la «verdadera» puntuación de la persona en el test (V), y otro, el error (e) que inevitablemente va asociado a todo proceso de medición; es decir, según el modelo:

$$X = V + e$$

A partir de este modelo y unas asunciones mínimas, la teoría clásica desarrollará todo un conjunto de deducciones encaminadas a estimar la cuantía del error que afecta a las puntuaciones de los test. Los inicios fueron como siempre inseguros, pues no era fácil abrirse paso en una psicología poco dada a veleidades cuantitativas, pero los nuevos enfoques se impusieron con rapidez y la mayoría de las universidades incluyeron cursos de teoría de los test en el currículum de sus licenciados. Cuenta Joncich (1968) en la biografía de E. L. Thorndike que cuando este envió una copia de su libro pionero sobre medición (Thorndike, 1904) a su antiguo profesor William James, incluyó una nota diciéndole que obligase a leerlo a todos sus investigadores, pero que no se le ocurriese abrirlo a él, pues las figuras, curvas y fórmulas que contenía le volverían loco. Ello puede dar una idea de la acogida que se esperaba por parte de la

© Ediciones Pirámide

psicología dominante de la época. Los años siguientes conocieron una actividad psicométrica frenética tanto en el campo de la teoría como en la construcción y tecnología de los test, así como en el campo del escalamiento psicofisico y psicológico (Thurstone, 1927a, 1928b; Thurstone y Chave, 1929), muy cercanos por entonces al ámbito de los test. Guilford (1936) tratará de sintetizar en su clásico manual *Psychometric Methods* lo fundamental de los tres campos, teoría de los test, escalamiento psicológico y psicofísico, algo que nadie volvería a intentar, salvo la reedición de 1954 de su propio libro (Guilford, 1954), pues se habían vuelto lo suficientemente complejos como para exigir cada uno de ellos un tratamiento aparte. Esos años de incubación y desarrollo dan lugar también a la institucionalización, y así, en 1936, se funda la sociedad psicométrica americana con Thurstone a la cabeza y cuyo órgano de expresión será la revista *Psychometrika.* Como complemento de carácter aplicado, frente al más teórico de *Psychometrika,* aparecerá unos años más tarde *Educational and Psychological Measurement* (1941); luego seguirían otras muchas entre las que cabe destacar *The British Journal of Mathematical and Statistical Psychology* (1948), *Journal of Educational Measurement* (1964), *Journal of Educational Statistics* (1976) y *Applied Psychological Measurement* (1977), por citar algunas.

En 1947 Thurstone publica su clásico texto *Análisis factorial múltiple,* técnica estadística con orígenes en el campo psicométrico (Burt, 1941; Kelley, 1928; Spearman, 1927), que aportará un notable avance para la construcción, análisis y validación de los test. En un campo como el del análisis factorial y las técnicas multivariadas en general, en el que se han producido avances tan gigantescos, potenciados por las facilidades informáticas y los refinamientos estadísticos actuales, resulta, sin embargo, refrescante releer el libro de Thurstone y admirar la cordura psicológica que lo impregna. Y es que cuando se piensa que un análisis factorial que un ordenador personal actual despacha en unos segundos podía llevar meses a un equipo de investigadores de entonces, se entiende que se afinasen las hipótesis y aguzase el ingenio antes de someterlas a prueba.

Por los años cuarenta también publicará Stevens (Stevens, 1946) su famoso trabajo sobre las escalas de medida, que obligará a los estudiosos de la teoría de los test a plantearse el estatus teórico de sus mediciones, además de sus propiedades empíricas, así como a terciar en la polémica que se abre entonces, y llega hasta nuestros días (Borsboom, 2005; Gaito, 1980; Michell, 1986; Townsend y Ashby, 1984), sobre las relaciones entre las escalas y sus implicaciones en el uso de las distintas técnicas estadísticas.

Pero la síntesis por antonomasia de la teoría clásica de los test será realizada por Gulliksen (1950) en su más que clásico libro *Theory of Mental Test,* que constituye, sin duda, la exposición mejor estructurada del corpus clásico. Gulliksen, antiguo estudiante, luego ayudante y colega de Thurstone, reconoce explícitamente el papel del maestro en su formación, y en especial de su libro *Fiabilidad y validez de los test* (Thurstone, 1931), ya agotado por las fechas en las que Gulliksen escribía el suyo. También en los años cincuenta aparecerán las primeras recomendaciones técnicas para el uso de los test (*Technical Recommendations for Psychological Test and Diagnostic Techniques,* 1954), que conocerán diversas actualizaciones y revisiones, la última en 2014.

También el escalamiento tendrá su clásico en los años cincuenta con el libro de Torgerson (1958) *Theory and Methods of Scaling.* Las dos ramas hermanas, teoría de los test y escalamiento, seguirán sus propios caminos, y aunque en esta como en otras divisiones hay algo de arbitrario, pues la mayoría de los modelos podrían generalizarse tanto a personas como a estímulos (Mosier, 1940, 1941), también es verdad que existían bastantes problemas específicos que justificaban la división.

Florecer parece haber sido el sino de los años sesenta, y la teoría de los test no iba a ser menos. Precisamente, en 1968 aparecerá el libro de Lord y Novick (1968) *Statistical Theories of Mental Test Scores,* que sintetiza y reanaliza críticamente todo lo hecho anteriormente en teoría clásica de los test, abriendo nuevas y prometedoras perspectivas. En el libro se incluye además el trabajo de Birnbaum sobre los modelos de rasgo latente, que abrirán una línea completamente nueva en la teoría de los test, conocida hoy como «teoría de respuesta a los ítems» (TRI). Este vuelco del modo de hacer clásico propiciado por la TRI va a oscurecer la mejora, al menos teórica, que prometía aportar al planteamiento clásico la teoría de la generalizabilidad (TG) propuesta

© Ediciones Pirámide

por Cronbach y colaboradores (Cronbach, Rajaratnam y Gleser, 1963; Gleser, Cronbach y Rajaratnam, 1965; Cronbach, Gleser, Nanda y Rajaratnam, 1972). La TG puede considerarse una extensión del modelo clásico, más que una alternativa. Mediante el uso masivo del análisis de varianza, la TG pretenderá analizar las fuentes del error de medida de un modo sistemático y desglosado, frente al tratamiento globalizado de la teoría clásica. Pero las aparatosas complicaciones introducidas en relación con las ventajas prácticas aportadas, unidas a la aparición en la escena psicométrica del enfoque alternativo de los modelos de TRI, relegarán la TG a un papel secundario en el campo de la teoría de los test.

No sería exacto decir que en el libro de Lord y Novick aparecen las primeras aportaciones sobre la TRI (véase Muñiz y Hambleton, 1992, para un estudio detallado del origen y desarrollo de estos modelos), pero la exposición y sistematización de Birnbaum, unidas a la plataforma publicitaria que supuso el libro, resultaron fundamentales para la rápida expansión de los modelos. De hecho ya hacía ocho años que Rasch había publicado su libro (Rasch, 1960) sobre el modelo logístico de un parámetro.

Sea como fuere, el libro de Lord y Novick, de áspera lectura y notación infernal, marca un antes y un después en la teoría de los test: terminaba una época, la clásica, y comenzaba otra nueva marcada por el predominio absoluto de la TRI. La nueva no negaba la anterior, aunque, como señalara Lord (1980), utilizará poco de ella para su formulación. Veamos a grandes rasgos cuáles eran las innovaciones que aportaba la TRI. La teoría clásica hallábase enfrentada con dos problemas de fondo importantes que no encontraban una solución satisfactoria en el marco clásico. Por un lado, la medición de las variables no era independiente del instrumento utilizado, algo así como si la longitud de los objetos dependiese del tipo de regleta. Por ejemplo, si la inteligencia de dos personas se mide con test distintos, los resultados de las mediciones no están en la misma escala; luego estrictamente no sabríamos cuál de las dos es más inteligente. Naturalmente, se había desarrollado todo un conjunto de soluciones técnicas para paliar el problema y poder equiparar las puntuaciones, pero se carecía de una solución digna de una medición aspirante al adjetivo de científica. En pocas palabras, las mediciones no eran invariantes respecto de los instrumentos de medida. Por otro lado, las propiedades de los instrumentos (test, ítems) dependían del tipo de personas utilizadas para establecerlas, por lo que en puridad no eran propiedades de los instrumentos de medida, sino de la interacción de estos con los objetos medidos. Por ejemplo, un ítem resultaría fácil o difícil en función de la muestra de personas utilizada. En suma, los instrumentos de medida no eran invariantes respecto de las personas evaluadas. Pues bien, los modelos de TRI permitirán dar una solución adecuada a esos dos problemas de fondo, y además aportarán todo un conjunto de avances tecnológicos complementarios para la construcción y análisis de los test.

Una década de investigación intensa sobre los distintos aspectos de los nuevos modelos, tanto a nivel teórico como aplicado, permitirá a Lord (1980) sintetizar en un libro hoy clásico los avances acumulados. El libro abre la década de los ochenta, que conocerá una expansión inusitada de la literatura psicométrica bajo la óptica de la TRI y revitalizará áreas que se encontraban atascadas, tales como los bancos de ítems, el sesgo o los test referidos al criterio, por citar solo algunas. Como señala Anastasi (1988) en la sexta edición de su incombustible *Psychological Testing,* la década de los ochenta ha sido un período de avances inusuales en teoría de los test, tanto respecto al progreso tecnológico como a la sofisticación teórica o la mejora de la responsabilidad profesional. Esta década vendrá marcada por una hegemonía clara de los modelos de TRI, como puede comprobarse echando una ojeada a los congresos de las sociedades psicométricas europea y americana, o a las revistas, y por la aparición de monografías independientes para cada una de las áreas ahora vigorizadas por la TRI. No obstante, no debe sacar el lector la impresión de que esta hegemonía de la TRI supone la muerte del enfoque clásico, ni mucho menos, pues la parsimonia y sencillez del modelo lineal clásico lo hacen apropiado en numerosas situaciones en las que la maquinaria pesada de la TRI no puede maniobrar con eficacia.

A la vista de este panorama general de la teoría de los test esbozado en las líneas precedentes, se comprenderá la limitada dimensión de lo que se pretende presentar en las páginas que siguen, cuyo objetivo no es otro que ofrecer al lector de lengua hispana una exposición clara y comprensiva de los aspectos fundamentales del enfoque clásico de los test y de la teoría de respuesta a los ítems, indican-

© Ediciones Pirámide

do en cada caso las fuentes adecuadas que permitan a quien lo desee una mayor profundización y preparación para comprender cabalmente los recientes avances de la teoría de los test, que no son pocos.

Nota histórica sobre los test

El desarrollo de la teoría de los test esbozado en el apartado anterior corre parejo con la evolución de los test concretos que van surgiendo, y como es natural, ambos desarrollos influyen el uno sobre el otro: los avances teóricos sobre los test y estos y sus problemas sobre los progresos teóricos. Bien es verdad que el avance no ha sido completamente sincronizado y en algunos momentos históricos una línea se ha adelantado por un tiempo a la otra, para luego volver a equilibrarse, como buenos vasos comunicantes que son. Así como el hacer suele preceder al pensar, los test como instrumentos se han anticipado a su fundamentación teórica. Los orígenes remotos de los primeros test podrían rastrearse según Du Bois (1970) allá por el año 3000 a.C., cuando los emperadores chinos evaluaban la competencia profesional de sus oficiales. Pero los orígenes más cercanos que darán lugar a los actuales test hay que ubicarlos en aquellas primeras pruebas sensomotoras utilizadas por Galton (1822-1911) en su famoso laboratorio antropométrico de Kensington. En 1884, durante la Exhibición Internacional sobre la Salud, que tuvo lugar en Londres, por la módica suma de tres peniques Galton medía a los visitantes todo un conjunto de índices antropométricos y sensomotores luego utilizados en sus estudios, como, por ejemplo, las dimensiones de la cabeza, estatura, longitud de los brazos extendidos, peso, fuerza de ambas manos, capacidad respiratoria, agudeza visual de ambos ojos, altura sentado, longitud del brazo, agudeza auditiva, tiempo de reacción visual y auditivo, precisión al dividir una línea en dos y tres partes iguales, error al estimar la apertura de un ángulo de 90 grados y otro de 60, etc. Hoy encontramos natural que al tratar de relacionar estas medidas con el funcionamiento intelectual no se encontrase conexión alguna, pero la hipótesis galtoniana de origen (Galton, 1883) tenía su lógica: si los datos sobre los que operamos han de ser filtrados por los sentidos, aquellos que contasen con sensores más finos dispondrían de un campo más amplio de actuación. También cabe a Galton el honor de ser el primero que aplicó la tecnología estadística para analizar los datos provenientes de sus test, labor que continuará Pearson. Como bien señala Boring (1950), si la década de los ochenta del siglo pasado viene marcada por Galton, la de los noventa vendrá por Cattell y la primera de este siglo por Binet. James McKeen Cattell (1860-1944) será el primero en utilizar el término «test mental» en su artículo «Mental test and measurements» publicado en la revista *Mind* en 1890, pero sus test, al igual que los de Galton, a quien por cierto admiraba, eran de carácter sensorial y motor fundamentalmente, y el análisis de los datos dejó clara la nula correlación entre este tipo de pruebas y el nivel intelectual de las personas (Wissler, 1901).

Será Binet (Binet y Simon, 1905a) quien dé un giro radical en la filosofía de los test al introducir en su escala tareas de carácter más cognoscitivo, encaminadas a evaluar aspectos como el juicio, la comprensión y el razonamiento, que según él constituían los componentes fundamentales del comportamiento inteligente. La puntuación de los niños en la escala de Binet y Simon se expresaba en términos de edad mental, que no era otra cosa que la edad cronológica de los niños que obtenían la misma puntuación media que el niño evaluado. Seguramente no es ajena al éxito alcanzado por la escala esta forma tan sencilla y comprensible para los no profesionales de expresar las puntuaciones de los niños. En la revisión de la escala que llevó a cabo Terman en la Universidad de Stanford, y que se conoce como la revisión Stanford-Binet (Terman, 1916), se utilizó por primera vez el cociente intelectual (CI) para expresar las puntuaciones. La idea era originaria de Stern, que en 1911 propuso dividir la edad mental (EM) entre la cronológica (EC), multiplicando por 100 para evitar los decimales:

$$CI = \frac{EM}{EC} \times 100$$

Esta fructífera veta de las escalas individuales de inteligencia abierta por Binet en 1905 y que se continúa hasta nuestros días había intuido, mejor que la hipótesis galtoniana, que si se desea evaluar el potencial intelectual hay que utilizar tareas cognoscitivamente complejas, que se asemejen de algún modo al tipo de cometidos intelectuales que se

© Ediciones Pirámide

pretende predecir. Puede decirse que el éxito de estas escalas para predecir el rendimiento escolar débese en gran medida al parecido de las tareas exigidas por ambos lados, escala y escuela. De hecho la causa próxima para que Binet pusiese manos a la obra de confeccionar su escala fue un encargo del Ministerio de Instrucción Pública para la detección y educación de los niños con deficiencias de inteligencia que asistían a las escuelas de París. Para una buena exposición de la escala, así como las sucesivas revisiones llevadas a cabo, véase Anastasi (1988).

El paso siguiente en el devenir histórico de los test vendrá marcado por la aparición de los test de inteligencia colectivos, propiciados por la necesidad del ejército estadounidense en 1917 de seleccionar y clasificar a los soldados que iban a tomar parte en la Primera Guerra Mundial. Un comité dirigido por Yerkes diseñó a partir de diverso material ya existente, especialmente de test inéditos de Otis, los hoy famosos test Alfa y Beta, el primero para la población general y el segundo para utilizar con analfabetos o reclutas sin dominio del inglés. Tras numerosas revisiones, estos test todavía siguen en uso. Debido a su éxito en el ejército, una vez finalizada la guerra la industria y el resto de las instituciones en general adoptaron en masa el uso de los test, conociéndose una expansión sin precedentes durante la próxima década, aunque no siempre en las mejores condiciones, debido por un lado a las limitaciones de los propios test, todavía un tanto rudimentarios, y por otro al uso de las puntuaciones más allá de lo que era razonable.

Con la experiencia acumulada iba quedando cada vez más claro que una puntuación global de inteligencia tal como la medían estos test no describía con suficiente precisión los diferentes aspectos del funcionamiento intelectual de las personas, y se imponía la evaluación de características más específicas cara a realizar pronósticos particulares más precisos. Si bien ello ya se venía haciendo de un modo más o menos sistemático, con la aparición de las hoy clásicas baterías de aptitudes habrá de esperar a que la técnica del análisis factorial dé sus frutos a partir de los años treinta y sobre todo cuarenta. Su producto más genuino serán las *aptitudes mentales primarias* de Thurstone (Thurstone, 1938; Thurstone y Thurstone, 1941), que conformaban lo que entonces se consideraban los componentes fundamentales del funcionamiento inteligente: comprensión verbal, fluidez verbal, aptitud numérica, aptitud espacial, memoria, rapidez perceptiva y razonamiento general. Dando más importancia a un factor general de inteligencia que articularía jerárquicamente otros factores de grupo (escuela inglesa), o reclamando un plano de igualdad para factores múltiples (escuela americana), el hecho central fue que el análisis factorial había permitido estructurar, no sin polémicas, la otrora genérica puntuación global de la inteligencia. En ocasiones los modelos alcanzaron grados de fragmentación rayanos en el desmenuzamiento, como en el caso de Guilford (1967), que propone nada menos que 120 rasgos intelectuales. Los distintos modelos darán lugar a numerosas baterías de test (PMA, DAT, GATB, TEA, etc.) de uso habitual actualmente. Excede por completo del cometido de estas líneas tratar de decir algo acerca del problema subyacente de la naturaleza de la inteligencia; la literatura al respecto es ciertamente abundante, y se aconseja al lector interesado la consulta de textos en español como los de Juan-Espinosa (1997) o Colom (1995, 2002).

A la vez que se producían los desarrollos citados en el campo de los test cognoscitivos, también los test de personalidad se beneficiaban de los avances técnicos que se iban produciendo, especialmente los derivados del análisis factorial y otras técnicas multivariadas afines. Suele citarse como origen próximo de los cuestionarios de personalidad de carácter psicométrico la hoja de datos personales utilizada por Woodworth en 1917 durante la Primera Guerra Mundial para la detección de neuróticos graves. En la actualidad, la sofisticación técnica en la construcción y análisis de los test de personalidad, que son legión (CEP, EPI, MMPI, 16PF, CPI, etc.), no se diferencia en nada de la utilizada con los test de aptitudes, si bien existen problemas específicos en unos y otros. Por su parte, el psiquiatra suizo Rorschach propone en 1921 su famoso test proyectivo de manchas de tinta, al que seguirán otros test proyectivos de muy distinto tipo de estímulos y tareas, aunque basados en la discutible asunción de la proyección, entre los que cabe citar el TAT, CAT, test de frustración de Rosenzweig, etc. Sin embargo, la técnica proyectiva que puede considerarse pionera es la asociación de palabras o test de asociación libre, descrita por Galton y utilizada incluso en el laboratorio por Wundt

© Ediciones Pirámide

y Cattell, aunque siguiendo las prácticas de Kraepelin y Jung pronto se asoció con la clínica, especialmente con la psicoanalítica.

Quede ahí esta somera nota histórica para ubicar al lector en el devenir de los test. Una excelente panorámica, así como la descripción y clasificación de los test más importantes, pueden verse en Anastasi y Urbina (1997). Para estudios históricos más detallados véanse Goodenough (1949), Du Bois (1970) o Geisinger y Usher-Tate (2016), y para una información enciclopédica y detallada sobre test concretos, véanse las sucesivas ediciones de los *Mental Mesurement Yearbooks* editados por Buros.

2. MODELO LINEAL CLÁSICO

Como ya se ha señalado en el apartado anterior, el modelo lineal clásico hunde sus raíces en los trabajos pioneros de Spearman (1904, 1907, 1913). En los años siguientes conoce un rápido desarrollo y Gulliksen (1950) recoge y sistematiza lo hecho hasta entonces. Lord y Novick (1968) llevarán a cabo una reformulación y análisis del modelo, a la vez que abren nuevas y fructíferas vías para la teoría de los test. Exposiciones menos técnicas del corpus clásico pueden consultarse en Guilford (1936, 1954), Magnuson (1967), Allen y Yen (1979), Thorndike (1982), Crocker y Algina (1986), y en español Yela (1987), Muñiz (2002), Martínez Arias et al. (2006) o Abad et al. (2011).

A continuación se exponen los aspectos fundamentales del modelo lineal clásico, su formulación, asunciones y deducciones más significativas. Aunque resulte obvio, no conviene perder de vista que el objetivo central del modelo es la estimación de los posibles errores cometidos cuando se utilizan los test para medir variables psicológicas. La necesidad de un modelo para estimar los errores de medida, en psicología como en cualquier otra ciencia, proviene del hecho elemental de que los errores no son observables directamente; lo que se obtiene directamente tras utilizar un instrumento de medida es el valor empírico «mezclado» con el error cuya cuantía se desea estimar. La lógica general para la estimación de los errores es común a todas las ciencias, pero la naturaleza de lo psicológico añadirá incluso algunas complicaciones adicionales.

TABLA 1.1
Formulación del modelo

Modelo:	$X = V + e$
Supuestos:	1. $V = E(X)$
	2. $\rho(v,e) = 0$
	3. $\rho(e_j,e_k) = 0$
Definición:	Dos test, j y k, se denominan paralelos si la varianza de los errores es la misma en ambos $[\sigma^2(e_j) = \sigma^2(e_k)]$ y también lo son las puntuaciones verdaderas de los sujetos ($V_j = V_k$).

El modelo establece que la puntuación empírica (X) que obtiene una persona en un test es igual a la suma de dos componentes: la puntuación verdadera (V) de la persona en ese test y el error de medida (e) cometido en la medición. Es razonable pensar que la puntuación empírica que obtiene una persona cuando se le aplica un test en un momento dado no coincida exactamente con su verdadera puntuación en ese test, pues en ese momento puntual la persona está afectada por múltiples factores de difícil control que inciden en su conducta. Una clasificación bastante exhaustiva de estas hipotéticas fuentes de error puede consultarse en Stanley (1971). Si estos factores perjudican a la persona, obtendrá una puntuación empírica más baja de la que verdaderamente le correspondería; si le benefician, la obtendrá superior. Ahora bien, cuando se pasa un test a una persona no hay manera de saber su puntuación verdadera, lo único que tenemos es su puntuación empírica, es decir, los puntos que saca en el test; su puntuación verdadera habrá que estimarla basándonos en los supuestos del modelo.

Supuesto 1

La puntuación verdadera (V) es la esperanza matemática de la empírica: $V = E(X)$, donde X es la variable aleatoria «puntuación empírica de la persona». Este primer supuesto constituye en realidad una definición de la puntuación verdadera. El lector poco familiarizado con el concepto de esperanza

© Ediciones Pirámide

matemática puede hacerse una idea imaginando que se aplicase un test «infinitas» veces a la misma persona. También debe imaginar que cada aplicación no afecta a las otras y que la persona no cambia en el curso de las aplicaciones. En estas condiciones, la puntuación verdadera de la persona en el test sería la media aritmética de las puntuaciones empíricas obtenidas en las «infinitas» aplicaciones. La puntuación verdadera es, por tanto, un concepto matemático. A partir de los valores de X (puntuaciones empíricas), y bajo ciertos supuestos que se irán viendo, la TCT permite hacer estimaciones probabilísticas razonables acerca del valor de las puntuaciones verdaderas (V). Conviene entenderlo bien, pues a menudo se ha hecho una conceptualización platónica de las puntuaciones verdaderas, considerándolas algo mágico y estático, propiedad de las personas y que determina su conducta. Del modelo no se sigue esta interpretación circular; la puntuación empírica en un test es una muestra de conducta que si reúne ciertos requisitos de medida, y bajo ciertos supuestos, permite hacer inferencias probabilísticas fundadas. De esto trata la teoría de los test.

Supuesto 2

Se asume que no existe correlación entre las puntuaciones verdaderas de las personas en un test y sus respectivos errores de medida: $\rho(v, e) = 0$. En principio no hay razón para pensar que el tamaño de los errores vaya sistemáticamente asociado al tamaño de las puntuaciones verdaderas.

Supuesto 3

Los errores de medida de las personas en un test no correlacionan con sus errores de medida en otro test distinto: $\rho(e_j, e_k) = 0$. Si se aplican correctamente los test, los errores serán aleatorios en cada ocasión, y a priori no existe razón para que covaríen sistemáticamente unos con otros.

Hay que señalar que ninguna de las asunciones del modelo es comprobable empíricamente de un modo directo tal como están expresadas; por tanto, aunque plausibles y sensatas a priori, habrá que hacer deducciones que sí se puedan contrastar y confirmen o falseen el modelo. Novick (1966) y Lord y Novick (1968) ofrecen formulaciones axiomáticas rigurosas del modelo a las que se remite al lector ávido de elegancia matemática.

Definición

Finalmente, se definen teóricamente los test paralelos, asumiéndose implícitamente que se pueden construir de hecho. De un modo menos formal que el señalado en su definición, podría decirse que dos test se consideran paralelos si miden lo mismo pero con diferentes ítems. Lord y Novick (1968) han desarrollado además otros tipos de paralelismo. Denominan test «tau equivalentes» a aquellos con puntuaciones verdaderas iguales para las personas en ambas formas, pero con varianzas error no necesariamente iguales. Test «esencialmente tau equivalentes» serían aquellos en los que la puntuación verdadera de cada persona en uno de ellos es igual a la del otro más una constante:

$$V_1 = V_2 + K$$

Los autores desarrollan las implicaciones y extensiones de estas redefiniciones de paralelismo para la TCT.

3. DEDUCCIONES INMEDIATAS DEL MODELO

A continuación se presentan algunas de las deducciones que se siguen directamente del modelo y que se utilizarán más adelante. Véase en el apéndice su obtención.

$$e = X - V \qquad [1.1]$$

El error de medida (e) se define, por tanto, como la diferencia entre la puntuación empírica (X) y la verdadera (V).

$$E(e) = 0 \qquad [1.2]$$

La esperanza matemática de los errores de medida es cero; luego son errores insesgados.

$$\mu_x = \mu_v \qquad [1.3]$$

© Ediciones Pirámide

La media de las puntuaciones empíricas es igual a la media de las verdaderas.

$$\text{cov}\,(V, e) = 0 \qquad [1.4]$$

Las puntuaciones verdaderas no covarían con los errores, lo cual es inmediato del supuesto 2 del modelo.

$$\text{cov}\,(X, V) = \text{var}\,(V) \qquad [1.5]$$

La covarianza entre las puntuaciones empíricas y las verdaderas es igual a la varianza de las verdaderas.

$$\text{cov}\,(X_j, X_k) = \text{cov}\,(V_j, V_k) \qquad [1.6]$$

La covarianza entre las puntuaciones empíricas de dos test es igual a la covarianza entre las verdaderas.

$$\text{var}\,(X) = \text{var}\,(V) + \text{var}\,(e) \qquad [1.7]$$

La varianza de las puntuaciones empíricas es igual a la varianza de las verdaderas más la de los errores.

$$\rho(X, e) = \sigma_e/\sigma_x \qquad [1.8]$$

La correlación entre las puntuaciones empíricas y los errores es igual al cociente entre la desviación típica de los errores y la de las empíricas.

$$\mu_1 = \mu_2 = \cdots = \mu_k \qquad [1.9]$$

$$\sigma^2(X_1) = \sigma^2(X_2) = \cdots = \sigma^2(X_k) \qquad [1.10]$$

$$\rho(X_1, X_2) = \rho(X_1, X_3) = \cdots = \rho(X_j, X_k) \qquad [1.11]$$

Es decir, para K test paralelos, las medias [1.9], las varianzas [1.10] y las intercorrelaciones entre ellos [1.11] son iguales.

Las ocho primeras deducciones son meras tautologías, no se pueden contrastar empíricamente, mientras que sí se pueden someter a prueba empírica las tres últimas. Nótese que el modelo y las deducciones están formulados para los valores paramétricos de la población. En términos generales —ya se irá matizando esta afirmación—, la teoría clásica de los test asume que los estadísticos obtenidos en muestras suficientemente amplias constituyen estimadores apropiados de los valores de la población. Cuanto más amplias sean las muestras, más pertinente será esta lógica. Lord y Novick (1968) proporcionan algunos estimadores más refinados para el caso de que las muestras no sean suficientemente amplias o no se cumplan estrictamente las condiciones de paralelismo.

© Ediciones Pirámide

2 Fiabilidad

Las mediciones psicológicas, como las de cualquier otra ciencia, han de ser fiables, es decir, han de estar libres de errores de medida. Un instrumento de medida, en nuestro caso un test o una escala, se considera fiable si las medidas que se hacen con él carecen de errores de medida, son consistentes. Una balanza es fiable si cada vez que pesamos el mismo objeto nos da el mismo resultado. Análogamente, un test será fiable si cada vez que se aplica a las mismas personas da el mismo resultado. La balanza lo tiene más fácil: el mismo objeto puede pesarse varias veces sin problema, pero los humanos cambian de una vez para otra, y a veces puede resultar problemático saber con seguridad si la inestabilidad observada en las mediciones se debe a la imprecisión del instrumento o a los cambios legítimos operados por las personas. Los errores de medida de los que se ocupa la fiabilidad son aquellos no sometidos a control e inevitables en todo proceso de medir, sea fisico, químico, biológico o psicológico. En muchas ocasiones las diferencias entre una medición y otra no dependen solo de estos errores, pudiendo explicarse además por los cambios operados en las personas, debidos a procesos madurativos, intervenciones o eventos de cualquier otro tipo. Incluso las inconsistencias pueden tener sentido en el marco en el que se lleva a cabo la medición. En estos casos la inestabilidad de las mediciones requiere una explicación y carece de sentido atribuirla a los errores aleatorios. La fiabilidad no trata ese tipo de «errores», que han de venir explicados por los modelos manejados. En cada situación el psicólogo tendrá que identificar las fuentes de error que afecten a las mediciones y no achacar, por ejemplo, a la baja fiabilidad de los instrumentos de medida lo que puede ser sencillamente variabilidad legítima de la variable medida. Nótese bien que el concepto de fiabilidad no se contradice en absoluto con la naturaleza cambiante de la conducta humana, como de un modo superficial se ha sugerido en algunas ocasiones; que cambie lo medido no anula la exigencia —todo lo contrario— de que los instrumentos de medida sean precisos. ¿Cómo medir adecuadamente el cambio sin instrumentos precisos? No conviene confundir la fiabilidad del instrumento de medida con la estabilidad o modificabilidad del constructo medido. La fiabilidad se refiere a la estabilidad de las mediciones cuando no existen razones teóricas ni empíricas para suponer que la variable a medir haya sido modificada diferencialmente para las personas, por lo que se asume su estabilidad, mientras no se demuestre lo contrario. Por ejemplo, parece sensato suponer que si se mide la inteligencia espacial de unas personas un día determinado y también se hace al día siguiente, su valor ha de ser básicamente el mismo, por lo que las posibles mínimas diferencias esperadas podrían atribuirse razonablemente a los errores aleatorios inherentes a todo acto de medir. Un test no sería fiable si cada día generase mediciones diversas de una variable que se supone estable. Ahora bien, lo que es válido para la inteligencia espacial no tiene por qué serlo para otras variables; por ejemplo, parece que la hora del día puede explicar gran parte de la variabilidad del tiempo de reacción de las personas, por lo que será en cada caso el psicólogo quien investigue las fuentes de error de las mediciones. No obstante, a nadie se le ocurrirá decir que los relojes que miden el tiempo de reacción en milisegundos no

© Ediciones Pirámide

son fiables por el hecho de que las medidas varíen a lo largo del día; la fiabilidad del instrumento no va unida a la estabilidad de la variable medida a lo largo del tiempo. Para un análisis detallado de la dialéctica fiabilidad/estabilidad de la conducta desde el punto de vista de la psicología clínica, véanse Silva (1989) y en general Cronbach y Furby (1970). Una interesante revisión sobre el problema de la fiabilidad en las investigaciones psicológicas puede verse en Schmidt y Hunter (1996).

En este capítulo se analizarán los distintos modos de estimar la fiabilidad de los test y la problemática implicada.

1. COEFICIENTE DE FIABILIDAD

El coeficiente de fiabilidad, $\rho_{XX'}$, se define como la correlación entre las puntuaciones obtenidas por las personas evaluadas en dos formas paralelas de un test, X y X'.

Es un indicador de la estabilidad de las medidas, pues si aplicamos un test X a una muestra de personas y pasado un tiempo aplicamos a las mismas personas una forma paralela X', dado que ambas formas miden lo mismo, si no hubiese errores aleatorios de medida, la correlación debería ser perfecta: $\rho_{XX'} = 1$. Por tanto, el grado en el que $\rho_{XX'}$ se aleja de 1 nos indicará en qué medida nuestras mediciones están afectadas por errores aleatorios de medida, siempre en el supuesto, claro está, de que las dos formas paralelas, X y X', realmente lo sean.

De la definición dada para el coeficiente de fiabilidad y de los supuestos del modelo se deriva fácilmente (véase apéndice):

$$\rho_{XX'} = \frac{\sigma_V^2}{\sigma_X^2} \qquad [2.1]$$

$$\rho_{XX'} = 1 - \frac{\sigma_e^2}{\sigma_X^2} \qquad [2.2]$$

A partir de estas dos fórmulas es imposible calcular empíricamente $\rho_{XX'}$, dado que el valor de $(\sigma_v)^2$ y $(\sigma_e)^2$ no se puede obtener de las respuestas de las personas a los ítems. No obstante, son útiles para dar una idea conceptual de lo que representa el coeficiente de fiabilidad. Por la primera, [2.1], se ve que $\rho_{XX'}$ indica la proporción que la varianza verdadera es de la empírica. Si no hubiese errores aleatorios, entonces $(\sigma_v)^2 = (\sigma_x)^2$ y $\rho_{XX'} = 1$. Tal vez se vea más claro todavía en [2.2]: si $(\sigma_e)^2 = 0$, $\rho_{XX'} = 1$; y si $(\sigma_e)^2 = (\sigma_X)^2$, o, lo que es lo mismo, $(\sigma_v)^2 = 0$, entonces $\rho_{XX'} = 0$. Se suele denominar índice de fiabilidad (ρ_{xv}) a la correlación entre las puntuaciones empíricas de un test y las verdaderas, siendo igual a la raíz cuadrada del coeficiente de fiabilidad (véase apéndice):

$$\rho_{XV} = \sqrt{\rho_{XX'}} = \frac{\sigma_V}{\sigma_X} \qquad [2.3]$$

Error típico de medida

Se denomina error típico de medida (σ_e) a la desviación típica de los errores de medida (e). Despejando de [2.2], su fórmula viene dada por:

$$\sigma_e = \sigma_X \sqrt{1 - \rho_{XX'}} \qquad [2.4]$$

2. ESTIMACIÓN EMPÍRICA DEL COEFICIENTE DE FIABILIDAD

Como ya se ha señalado, las fórmulas del coeficiente de fiabilidad expuestas hasta ahora no permiten calcular su valor empírico para una muestra determinada de personas. Para poder hacerlo hay que valerse de su definición: correlación entre las puntuaciones en dos formas paralelas.

Se tratará, en suma, de:

1. Elaborar las dos formas paralelas.
2. Aplicarlas a una muestra amplia de personas representativas de la población en la que se va a utilizar el test.
3. Calcular la correlación entre las puntuaciones de las personas en ambas formas.

Dicha correlación será precisamente el coeficiente de fiabilidad. Este método se denomina por razones obvias *método de las formas paralelas,* y es el que emana genuina y directamente del modelo. No es infrecuente denominar coeficiente de equivalencia al valor obtenido, aludiendo a que cierta-

© Ediciones Pirámide

mente indicaría el grado en el que ambas formas son equivalentes. Se suelen utilizar además otros dos métodos, denominados, respectivamente, «test-retest» y «dos mitades». Veamos en qué consisten.

Test-retest. Para calcular el coeficiente de fiabilidad por este método se aplica el mismo test en dos ocasiones a las mismas personas; la correlación entre las puntuaciones de las dos aplicaciones será el coeficiente de fiabilidad. Dado que obviamente un test es paralelo a sí mismo, este método es perfectamente congruente con el modelo, denominándose a la estimación obtenida «coeficiente de estabilidad», pues indica en qué grado son estables las mediciones realizadas en la primera aplicación del test.

Dos mitades. Por este método se aplica el test una sola vez, obteniéndose para cada persona las puntuaciones correspondientes a cada una de las mitades en las que se divide el test. El coeficiente de fiabilidad viene dado por la correlación entre esas dos mitades (que será la estimación de la fiabilidad del test mitad) *más una corrección* para obtener la fiabilidad del test total (esta corrección se verá más adelante cuando se exponga la fórmula de Spearman-Brown). La estimación así obtenida, más que equivalencia o estabilidad, como en los casos anteriores, indica la covariación o consistencia interna de las dos mitades; es, pues, un indicador de la consistencia interna del test.

Si bien la lógica de estos tres métodos es clara, su realización empírica plantea diversos problemas experimentales relativos a la validez interna, para los cuales el modelo no da especificaciones concretas, quedando al criterio del psicólogo para cada situación nueva planteada. A continuación se comentan algunos de ellos.

En el método de las formas paralelas el problema fundamental es la construcción de dichas formas paralelas. Es difícil a nivel teórico hacer un test que mida exactamente lo mismo que otro, pero con distintos ítems; tal vez, incluso, filosóficamente imposible, y en la práctica es enormemente laborioso. Si se superan los problemas y se dispone de dos (o más) formas paralelas, probablemente es el método más recomendable.

En el método test-retest una cuestión de difícil solución es delimitar el tiempo óptimo que debe transcurrir entre ambas aplicaciones. Si se deja mucho, se introduce una gran fuente de invalidez interna, a saber, la ignota influencia diferencial de ese período de tiempo en las personas; pero si transcurre poco tiempo, la invalidez interna se cuela vía memoria de lo realizado previamente. No hay regla universal: depende en gran parte del tipo de test, ya que es evidente que hay unos test más propensos a ser recordados que otros. Una aproximación rigurosa al problema general del test-retest puede consultarse en Jöreskog y Sörbom (1976).

Finalmente, el método de las dos mitades es muy funcional, pues solo exige una sola aplicación del test. No obstante, hay que garantizar que las mitades del test sean paralelas. No es recomendable, por ejemplo, considerar mitades la primera parte del test por un lado y la segunda por otro, pues las personas evaluadas llegarán más cansadas a la segunda; además, en muchos test cognoscitivos los ítems van aumentando en dificultad, por lo que la segunda parte resultaría más difícil que la primera. Para evitar esto es frecuente tomar como una mitad los ítems pares y como otra los impares, o usar algún otro tipo de apareamiento de los ítems. En definitiva, es un problema de control experimental.

Un factor del test a tener en cuenta para elegir un método u otro de los comentados, o de otros que se verán más adelante, es si se trata de un test de velocidad o de un test de potencia. Suele entenderse por test de velocidad aquel cuya realización no conlleva dificultad alguna, o, más exactamente, el que todas las personas son capaces de realizar, aunque difieran en la rapidez de ejecución. Por el contrario, un test de potencia o poder sería aquel en el que las diferencias entre las personas son generadas por su distinta capacidad intelectual para resolver las tareas propuestas. No hacen falta muchas explicaciones para entender que en la práctica la mayoría de los test suelen ser mixtos, variando la proporción de ambos componentes: en unos predomina más la velocidad, y en otros, la potencia.

Índices de velocidad-potencia

El grado de velocidad de un test influye en sus parámetros más importantes, como su fiabilidad, su validez, la estructura factorial de los ítems, o en su caso de la batería; de ahí que se hayan propuesto

© Ediciones Pirámide

diferentes índices para expresar la proporción velocidad/potencia (véase, por ejemplo, Donlon, 1978, para un buen análisis). Gulliksen (1950) sugiere el cociente entre la varianza de los errores cometidos y la varianza de los fallos (errores más no intentados). Cuanto más bajo sea el cociente, más de velocidad será el test; con el límite cero indicando que todos los fallos se deben a no-intentos, el test sería de velocidad pura.

$$IV = \frac{\sigma_e^2}{\sigma_F^2}$$

Lord y Novick (1968) proponen un índice equivalente al de Cronbach y Warrington (1951) que indica la proporción de varianza atribuible a la velocidad:

$$IV = 1 - \frac{\rho_{VP}^2}{\rho_{VV'}\rho_{PP'}} \qquad [2.5a]$$

donde:

- IV: Índice de velocidad.
- ρ_{VP}: Correlación entre el test administrado en condiciones de velocidad (V), esto es, con tiempo limitado para su ejecución, y el test administrado con tiempo «ilimitado», en condiciones de potencia (P).
- $\rho_{VV'}$: Coeficiente de fiabilidad del test en condiciones de velocidad.
- $\rho_{PP'}$: Coeficiente de fiabilidad del test en condiciones de potencia.

El mayor inconveniente práctico de este índice es que requiere aplicar el test dos veces, lo que no siempre es fácil.

Ejemplo

El test de inteligencia general BLSIV se aplicó a una muestra de 500 universitarios con un tiempo de cinco minutos. A la semana siguiente se les aplicó sin tiempo límite. La correlación entre las puntuaciones de los universitarios en ambas aplicaciones fue 0,60; la fiabilidad del test con tiempo limitado, 0,70, y sin tiempo límite, 0,80. Calcular el índice de velocidad del test.

$$IV = 1 - \frac{(0{,}60)^2}{(0{,}70)(0{,}80)} = 0{,}36$$

El 36% de la varianza de las puntuaciones sería atribuible a la velocidad de respuesta de los universitarios.

Otro indicador más sencillo y de fácil uso es el cociente de velocidad propuesto por Stafford (1971):

$$CV = \frac{\sum NI}{\sum E + \sum O + \sum NI} \times 100 \qquad [2.5b]$$

donde:

- E: Número de errores de cada persona.
- O: Número de omisiones de cada persona.
- NI: Número de ítems no intentados por cada persona.

Cuando un test es de velocidad pura, ΣE y ΣO son cero, y en consecuencia $CV = 100$. En el caso de un test de potencia pura, $\Sigma NI = 0$, por lo que $CV = 0$, o, lo que es lo mismo, el cociente de potencia, que es el complementario del de velocidad, será 100.

Suprimiendo el sumatorio de la fórmula anterior, se tiene el cociente de velocidad para cada persona. Asimismo, si en vez de los datos correspondientes a una persona utilizamos los de un ítem, tendremos el cociente de velocidad del ítem.

3. ESTIMACIÓN DE LAS PUNTUACIONES VERDADERAS

Conocida la fiabilidad del test por alguno de los métodos expuestos, se pueden hacer ciertas estimaciones acerca de las puntuaciones verdaderas de las personas en el test, o, lo que es lo mismo, se pueden hacer estimaciones acerca de la cantidad de error que afecta a las puntuaciones empíricas. A continuación se exponen las tres estrategias más comúnmente utilizadas por la psicometría clásica, advir-

© Ediciones Pirámide

tiendo desde el principio que estas estimaciones han de tomarse con extremada cautela cuando se hacen para una persona en particular, siendo más apropiadas para la descripción de grupos.

a) *Estimación mediante la desigualdad Chebychev*

Como es sabido, la desigualdad de Chebychev establece que para toda variable X con media $\overline{X}$ y desviación típica S_X:

$$\forall K \quad P\{|X - \overline{X}| \leq K(S_X)\} \geq 1 - \frac{1}{K^2} \quad [2.6]$$

que traducido a la terminología psicométrica del modelo clásico:

$$\forall K \quad P\{|X - V| \leq K(\sigma_e)\} \geq 1 - \frac{1}{K^2} \quad [2.7]$$

Veamos el paso de la fórmula general [2.6] a la terminología psicométrica:

Se ha sustituido la media $\overline{X}$ por V, puesto que en el modelo clásico $E(X) = \mu_X = V$. Por su parte, la desviación típica S_X se ha sustituido por σ_e, puesto que en el modelo la varianza de X para un valor dado de V (persona o clase de personas) es igual a la varianza de los errores $(\sigma_e)^2$. Nótese que al fijar V lo que varían las empíricas se debe únicamente a la variación de los errores. Asimismo, la varianza de los errores para un determinado valor de V (persona o clase de personas) es igual a la varianza de los errores de la población σ_e^2 dado que $\rho_{ve} = 0$. En definitiva:

$$\sigma^2(X|V) = \sigma^2(e|V) = \sigma_e^2$$

Ejemplo

Se aplicó un test de rapidez perceptiva a una muestra representativa de 1.000 personas, obteniéndose una media de 70 puntos, una desviación típica de 10 y un coeficiente de fiabilidad de 0,64. Al nivel de confianza del 99%, ¿qué puntuación verdadera se estimará a las personas que obtuvieron una puntuación empírica en el test de 80 puntos?

Datos:

$N = 1.000;\ S_X = 10;\ r_{XX'} = 0{,}64;\ X = 80$

$$S_e = S_X\sqrt{1 - r_{XX'}} = 10\sqrt{1 - 0{,}64} = 6$$

$1 - 1/K^2 = 0{,}99$; por tanto: $K = 10$.
Sustituyendo en [2.7]:

$$P\{|80 - V| \leq (10)(6)\} \geq 0{,}99$$
$$P\{|V - 80| \leq (10)(6)\} \geq 0{,}99$$
$$P\{-60 \leq V - 80 \leq 60\} \geq 0{,}99$$
$$P\{20 \leq V \leq 140\} \geq 0{,}99$$

Al nivel de confianza del 99% estimamos que la puntuación verdadera en el test de rapidez perceptiva para las personas que obtuvieron una empírica de 80 estará ¡entre 20 y 140! Ciertamente es una pobre estimación; el intervalo es demasiado amplio, y ello se debe a dos razones: una, que el coeficiente de fiabilidad es bajo (0,64), y otra, que este método de estimación paga un caro tributo al no hacer ninguna asunción-restricción sobre la forma de la distribución de X, aunque, eso sí, es válido sea cual sea esta, lo cual es interesante, pero a un precio prohibitivo a base de abrir el intervalo confidencial.

b) *Estimación basada en la distribución normal de los errores*

Este es el método más utilizado en los textos clásicos. Una forma de evitar una estimación tan genérica como la anterior es asumir que los errores de medida, y por ende las puntuaciones empíricas, para un valor dado de V (persona o clase de personas con la misma puntuación verdadera) se distribuyen según la curva normal:

$$f(e|V) \sim N(0, \sigma_e^2)$$

Dado que $X = V + e$, con V constante, es inmediato que

$$f(X|V) \sim N(V, \sigma_e^2)$$

© Ediciones Pirámide

Adviértase que el precio a pagar por reducir la amplitud del intervalo es la asunción de normalidad que se añade al modelo. Hasta ahora no se había hecho ninguna asunción sobre la forma de las distribuciones de las puntuaciones. Esta asunción de normalidad e igualdad de las varianzas condicionales a lo largo de la escala de las puntuaciones (homoscedasticidad) ha sido cuestionada con frecuencia, especialmente en lo concerniente a los valores extremos de la escala. Si no se cumple, lo cual es bastante probable en la práctica, no sería muy preciso utilizar el mismo error típico de medida (ETM) para todas las personas, independientemente de su puntuación en el test, como se hace habitualmente. Según los datos de Feldt, Steffan y Gupta (1985), el valor máximo del ETM correspondería a la franja de puntuaciones medias, encontrando grandes diferencias para los distintos niveles de puntuaciones. Este problema no hallará una respuesta apropiada en el marco de la teoría clásica de los test, y habrá que esperar al desarrollo de los modelos de teoría de respuesta a los ítems para una solución adecuada. Mediante la función de información del test, estos modelos permitirán estimar el error de medida para los distintos niveles de competencia de las personas.

Supuesto esto, para estimar la puntuación verdadera en el test se establece el intervalo confidencial en torno a la puntuación empírica del modo habitual. Veámoslo para el ejemplo anterior y comparemos aquel intervalo con este:

1. Al NC del 99% corresponde una Z_C de ±2,58.
2. Error máximo admisible: $(Z_C)(\sigma_e) = = (2{,}58)\,(6) = 15{,}48$.
3. Intervalo confidencial:

$$(80 - 15{,}48) \leqslant V \leqslant (80 + 15{,}48)$$
$$(64{,}52 \leqslant V \leqslant 95{,}48)$$

Obsérvese la reducción del intervalo confidencial respecto al método anterior para los mismos datos, aunque sigue siendo excesivamente amplio debido a la baja fiabilidad del test.

c) Estimación según el modelo de regresión

Como es bien sabido, en el modelo de regresión lineal el pronóstico de una variable Y a partir de otra X, según el criterio de mínimos cuadrados, viene dado por la expresión:

$$Y' = \rho_{XY}\left(\frac{\sigma_Y}{\sigma_X}\right)(X - \bar{X}) + \bar{Y} \qquad [2.8]$$

Dado que ese es nuestro problema, estimar V a partir de X, traduzcamos esta expresión a nuestra terminología, donde Y, lo que se desea pronosticar, pasa a ser V, y X sigue siendo X:

$$V' = \rho_{XV}\left(\frac{\sigma_V}{\sigma_X}\right)(X - \bar{X}) + \bar{V}$$

Ahora bien, como hemos visto, $\bar{V} = \bar{X}$ y $\sigma_V/\sigma_X = \rho_{XV}$; luego:

$$V' = \rho^2_{XV}(X - \bar{X}) + \bar{X}$$

y como $\rho^2_{XV} = \rho_{XX'}$

$$V' = \rho_{XX'}(X - \bar{X}) + \bar{X} \qquad [2.9]$$

Mediante esta fórmula podemos hacer estimaciones puntuales de V a partir de X, conociendo el coeficiente de fiabilidad, la media del test y la puntuación empírica.

Ejemplo

Se aplicó un test de comprensión verbal a una muestra de 500 personas y se obtuvo una media de 40 y un coeficiente de fiabilidad de 0,80. ¿Qué puntuación verdadera en el test se pronosticará a las personas que obtuvieron una puntuación empírica de 60 puntos?

Sustituyendo en [2.9]:

$$V' = 0{,}80(60 - 40) + 40 = 56$$

Es decir, a las personas con una puntuación empírica de 60 se les pronostica una verdadera de 56. Ahora bien, el modelo de regresión utilizado lo único que garantiza es que «a la larga» los errores de

© Ediciones Pirámide

pronóstico cometidos son mínimos, según el criterio de mínimos cuadrados, pero la puntuación pronosticada V' no siempre coincidirá con V, denominándose precisamente a esa diferencia error de estimación. Es por ello por lo que para asegurarse de los pronósticos, en vez de realizar estimaciones puntuales, se establecen intervalos confidenciales en torno a la puntuación pronosticada V'. Para establecer dichos intervalos nos valemos del error típico de estimación, que es la desviación típica de los errores de estimación, y que en su forma general, para dos variables X e Y, viene dado por

$$\sigma_{Y \cdot X} = \sigma_Y \sqrt{1 - \rho_{XY}^2} \qquad [2.10]$$

Que traducido a la terminología y supuestos del modelo lineal clásico (véase apéndice) puede expresarse así:

$$\sigma_{V \cdot X} = \sigma_X \sqrt{1 - \rho_{XX'}} \sqrt{\rho_{XX'}} \qquad [2.11]$$

o también, teniendo en cuenta que $\sigma_X \sqrt{1 - \rho_{XX'}} = \sigma_e$:

$$\sigma_{V \cdot X} = \sigma_e \sqrt{\rho_{XX'}} \qquad [2.12]$$

Asumiendo que los errores de estimación se distribuyen normalmente en torno a V', se pueden establecer los correspondientes intervalos confidenciales. Establezcamos dicho intervalo para el mismo ejemplo de los apartados anteriores *a*) y *b*) y comparemos los resultados.

Recuérdese que los datos eran: NC: 99 %; $N = 1.000$; $S_X = 10$; $r_{XX'} = 0{,}64$; $X = 80$; $\bar{X} = 70$.

1. Al NC del 99 % corresponde una Z_C de ±2,58.
2. $\sigma_{V \cdot X} = \sigma_X \sqrt{1 - \rho_{XX'}} \sqrt{\rho_{XX'}} =$
   $= 10\sqrt{1 - 0{,}64}\sqrt{0{,}64} = 4{,}8$.
3. Error máximo:
   $$(Z_C)(\sigma_{V \cdot X}) = (2{,}58)(4{,}8) = 12{,}384.$$
4. $V' = \rho_{XX'}(X - \bar{X}) + \bar{X} = 0{,}64(80 - 70) + 70 =$
   $= 76{,}4$.
5. Intervalo confidencial: V' ± error máximo:
   $$(64{,}016 \leqslant V \leqslant 88{,}784)$$

Nótese que el intervalo así obtenido siempre será menor o igual al obtenido en el apartado *b*). Allí se utilizaba σ_e para establecerlo, mientras que aquí se utiliza $\sigma_{V \cdot X}$, que es igual a $\sigma_e \sqrt{\rho_{XX'}}$. Cuando $\rho_{XX'}$ tomase su valor máximo de 1 (por otra parte inalcanzable empíricamente), entonces $\sigma_{V \cdot X'} = \sigma_e$, y los intervalos serían iguales.

Nota. Todas las fórmulas utilizadas en este apartado están expresadas en puntuaciones directas. A continuación se ofrecen para la escala diferencial y típica. Se propone como ejercicio al lector la obtención de estas sencillas transformaciones de escala.

Directas	$Y' = \rho_{XY}\left(\frac{\sigma_Y}{\sigma_X}\right)(X - \bar{X}) + \bar{Y}$
diferenciales	$y' = \rho_{XY}(\sigma_Y/\sigma_X)x$
típicas	$Z_{Y'} = \rho_{XY} Z_X$
Directas	$V' = \rho_{XX'}(X - \bar{X}) + \bar{X}$
diferenciales	$v' = \rho_{XX'}x$
típicas	$Z_{Y'} = \sqrt{\rho_{XX'}} Z_X$
Directas	$\sigma_{Y \cdot X} = \sigma_Y \sqrt{1 - \rho_{XY}^2}$
diferenciales	Igual que en directas
típicas	$\sigma_{Z_x \cdot Z_Y} = \sqrt{1 - \rho_{XY}^2}$
Directas	$\sigma_{Y \cdot X} = \sigma_X \sqrt{1 - \rho_{XX'}} \sqrt{\rho_{XX'}}$
diferenciales	Igual que en directas
típicas	$\sigma_{Z_v \cdot Z_X} = \sqrt{1 - \rho_{XX'}} \sqrt{\rho_{XX'}}$

4. FIABILIDAD DE LAS DIFERENCIAS

Existen numerosas situaciones aplicadas en psicología y educación en las que interesa estudiar las diferencias existentes entre las puntuaciones de las personas en un test y sus puntuaciones en otro; piénsese, por ejemplo, en la evaluación de intervenciones, terapias, tratamientos, etc. Para poder interpretar las diferencias encontradas es imprescindible disponer

© Ediciones Pirámide

de alguna medida de su fiabilidad. Dos diferencias iguales pueden tener muy distinto valor científico para el psicólogo en función de su fiabilidad.

Para dos test X y Z la fiabilidad de las diferencias entre sus puntuaciones: $(X - Z) = d$, como fácilmente se puede derivar (véase apéndice), viene dada por:

$$\rho_{dd'} = \frac{\sigma_X^2 \rho_{XX'} + \sigma_Z^2 \rho_{ZZ'} - 2\sigma_X \sigma_Z \rho_{XZ}}{\sigma_X^2 + \sigma_Z^2 - 2\sigma_X \sigma_Z \rho_{XZ}} \quad [2.13]$$

donde:

$\rho_{dd'}$: Coeficiente de fiabilidad de las diferencias.
σ_X^2: Varianza de las puntuaciones del test X.
σ_Z^2: Varianza de las puntuaciones del test Z.
$\rho_{XX'}$: Coeficiente de fiabilidad del test X.
$\rho_{ZZ'}$: Coeficiente de fiabilidad del test Z.
$\rho_{XZ'}$: Correlación entre ambos test.

Si ambos test se expresan en la misma escala para una mejor interpretabilidad de las diferencias, clásicamente en la escala de típicas, aunque cualquier otra es posible, en cuyo caso sus varianzas serían iguales ($\sigma_X^2 = \sigma_Z^2$), la fórmula anterior se simplifica:

$$\rho_{dd'} = \frac{\rho_{XX'} + \rho_{ZZ'} - 2\rho_{XZ}}{2(1 - \rho_{XZ})} \quad [2.14]$$

El error típico de medida de las diferencias (σ_{ed}), análogamente a lo visto para un solo test, vendrá dado por:

$$\sigma_{ed} = \sigma_d \sqrt{1 - \rho_{dd'}} \quad [2.15]$$

donde:

σ_d: Desviación típica de las diferencias.
$\rho_{dd'}$: Coeficiente de fiabilidad de las diferencias.

Ejemplo

El coeficiente de fiabilidad de un test de inteligencia espacial fue 0,60, y el de otro test también de inteligencia espacial, 0,50. La correlación obtenida entre las puntuaciones de 1.000 personas en ambos test fue de 0,40. ¿Cuál será el coeficiente de fiabilidad de las diferencias entre las puntuaciones de las personas en ambos test?

$$r_{dd'} = \frac{0{,}60 + 0{,}50 - 2(0{,}40)}{2(1 - 0{,}40)} = 0{,}25$$

La fiabilidad es muy baja. En una situación de este tipo más vale abstenerse de emitir juicios acerca de las diferencias halladas, pues son poco fiables.

Este tipo de análisis de las diferencias entre las puntuaciones de dos test (o más) es muy usual en el análisis de perfiles psicológicos. Como es bien sabido, un perfil psicológico no es otra cosa que la representación, generalmente gráfica, de las puntuaciones de una persona o de un grupo en determinadas variables, expresadas todas ellas en la misma escala (misma media y misma desviación típica) para percatarse mejor del sentido de las diferencias entre las variables.

El coeficiente de fiabilidad de las diferencias es, asimismo, fundamental para la medida del cambio experimentado por las puntuaciones de las personas en alguna variable psicológica o educativa. Casos típicos se presentan cuando hay que evaluar el impacto de terapias clínicas, programas de aprendizaje o cualquier otro tipo de intervención.

Conviene señalar finalmente que, en contra de la extendida costumbre, no está justificado hacer comparaciones individuales entre las puntuaciones empíricas de dos personas en un test, de una persona en dos test, etc., basándose en los errores típicos de medida que se pueden derivar (véase apartado siguiente). Este tipo de inferencias, como bien señalan Lord y Novick (1968), son incorrectas, y solo quedarían justificadas si los pares de personas que se desea comparar se cogiesen al azar y sin referencia a su puntuación empírica; si así se hiciese, a la larga (estrictamente infinito) las inferencias tendrían sentido globalmente, pero nunca para dos personas específicas elegidas, porque interesa comparar precisamente sus puntuaciones empíricas.

5. TIPOS DE ERRORES DE MEDIDA

Hasta aquí se han definido dos tipos de errores: el error de medida y el error de estimación. Cabe citar además (Gulliksen, 1950; Lord y Novick, 1968) el

© Ediciones Pirámide

error de sustitución y el error de predicción. A continuación se exponen los cuatro junto con sus respectivas desviaciones típicas y se comentan brevemente.

1. Error de medida

$$e = X - V$$
$$\sigma_e = \sigma_X \sqrt{1 - \rho_{XX'}} \qquad [2.4]$$

2. Error de estimación

$$e = V - V'$$
$$\sigma_{V \cdot X} = \sigma_X \sqrt{1 - \rho_{XX'}} \sqrt{\rho_{XX'}} \qquad [2.11]$$

3. Error de sustitución

$$e = X_1 - X_2$$
$$\sigma_{e(s)} = \sigma_X \sqrt{1 - \rho_{XX'}} \sqrt{2} \qquad [2.16]$$

4. Error de predicción

$$e = X_1 - X_1'$$
$$\sigma_{e(P)} = \sigma_X \sqrt{1 - \rho_{XX'}} \sqrt{1 + \rho_{XX'}} \qquad [2.17]$$

El error de medida y el error de estimación son, respectivamente, como ya se ha visto, la diferencia entre la puntuación empírica y la verdadera $(X - V)$, y la diferencia entre la puntuación verdadera y la verdadera pronosticada $(V - V')$. Sus desviaciones típicas, también previamente definidas, se denominan «error típico de medida» (σ_e) y «error típico de estimación» $(\sigma_{V \cdot X})$, respectivamente.

El error de sustitución es la diferencia entre las puntuaciones en un test (X_1) y las obtenidas en otro paralelo (X_2); es, por tanto, el error de medida que se generaría al sustituir una medición por otra proveniente de un test paralelo. Su desviación típica dada por [2.16] es el error típico de las diferencias entre dos test paralelos.

El error de predicción es la diferencia entre las puntuaciones en un test (X_1) y las puntuaciones pronosticadas en ese test (X_1') a partir de una forma paralela X_2. Es el error que se cometería al utilizar, en vez de las mediciones de un test, aquellas pronosticadas en ese test a partir de una forma paralela. Es decir, X_1' son los pronósticos realizados mediante la recta de regresión de X_1 sobre X_2, que viene dada según el modelo general [2.8] adaptado a nuestra terminología por:

$$X_1' = \rho_{12} \frac{\sigma_1}{\sigma_2} (X_2 - \bar{X}_2) + \bar{X}_1 \qquad [2.18]$$

A modo de ejercicio, trate el lector de derivar los cuatro errores típicos partiendo de las varianzas de los correspondientes errores de medida. Trate asimismo de ordenarlos de menor a mayor.

6. FACTORES QUE AFECTAN A LA FIABILIDAD

6.1. Fiabilidad y variabilidad

La fiabilidad de un test no depende únicamente de sus características propias, sino también del tipo de muestra de personas utilizadas para calcularla, lo cual constituye una seria limitación para el modelo clásico, pues se está describiendo un instrumento de medida, como es el test, en función de los «objetos» medidos, las personas. Uno de los aspectos de la muestra que influye en la fiabilidad de su variabilidad, el coeficiente de fiabilidad, aumenta al aumentar la variabilidad de la muestra. La razón es bien simple: el coeficiente de fiabilidad se ha definido como la correlación entre dos formas paralelas de un test, y es bien sabido que la correlación viene afectada por la variabilidad del grupo, aumentando con esta. Por tanto, un dato imprescindible para la interpretación del coeficiente de fiabilidad es la variabilidad de la muestra en la que se calculó; en otras palabras, un test no tiene un coeficiente de fiabilidad fijo, este depende de la variabilidad de la muestra en la que se calcule.

En suma, al aumentar la variabilidad de la muestra, aumenta el valor del coeficiente de fiabilidad, proponiéndose en este apartado una fórmula que permite estimar ese aumento. La fórmula es apropiada si se cumple el supuesto en el que se basa, a saber, que la varianza de los errores de medida en el test es la misma en ambas poblaciones, la menos variable y la más variable.

Según [2.4] el error típico de medida venía dado por:

$$\sigma_e = \sigma_X \sqrt{1 - \rho_{XX'}}$$

© Ediciones Pirámide

El supuesto citado puede expresarse, por tanto: $(\sigma_{e1})^2 = (\sigma_{e2})^2$, o, más explícitamente, sustituyendo $(\sigma_e)^2$ por su valor en ambas poblaciones:

$$\sigma_1^2(1 - \rho_{11'}) = \sigma_2^2(1 - \rho_{22'})$$

Despejando $\sigma_{22'}$:

$$\rho_{22'} = 1 - \frac{\sigma_1^2}{\sigma_2^2}(1 - \rho_{11'}) \qquad [2.19]$$

donde:

$\rho_{11'}$: Coeficiente de fiabilidad en la población 1.
$\rho_{22'}$: Coeficiente de fiabilidad en la población 2.
σ_1^2: Varianza empírica en la población 1.
σ_2^2: Varianza empírica en la población 2.

Ejemplo

El coeficiente de fiabilidad de un test de coordinación visomotora en una muestra de universitarios que habían superado las pruebas de selectividad fue de 0,64, obteniéndose una varianza de 25. ¿Cuál habría sido el coeficiente de fiabilidad del test si se hubiese calculado con los datos de todos los aspirantes cuya varianza fue 100, y no solo utilizando los que superaron la selectividad?

$$\rho_{22'} = 1 - \frac{25}{100}(1 - 0{,}64) = 0{,}91$$

Queda patente con el ejemplo la relatividad del coeficiente de fiabilidad, que pasa de 0,64 a 0,91 al aumentar la variabilidad de 25 a 100.

Adviértase que la pertinencia de la fórmula [2.19] puede comprobarse empíricamente aplicando el test a las dos poblaciones de interés y luego comparando los resultados hallados con los estimados por la fórmula. Como Lord y Novick (1968) señalaran, esta fórmula debe usarse con precaución, pues el fuerte supuesto en el que se basa no siempre se cumple. Si, por ejemplo, ocurriese que σ_{e1} fuese menor que σ_{e2} y utilizásemos la fórmula, se sobreestimaría $\rho_{22'}$, ya que, si $\sigma_{e1} < \sigma_{e2}$, entonces

$$\sigma_1\sqrt{1 - \rho_{11'}} < \sigma_2\sqrt{1 - \rho_{22'}}$$

con lo que

$$\rho_{22'} < 1 - \frac{\sigma_1^2}{\sigma_2^2}(1 - \rho_{11'})$$

6.2. Fiabilidad y longitud

La fiabilidad de un test también depende de su longitud, entendiendo por longitud el número de ítems que contiene. En principio parece enjundioso pensar que cuantos más ítems se utilicen para evaluar una variable, mejor podremos muestrear los diferentes aspectos que la conforman y más fiable será la medida obtenida. En el límite, infinitos ítems, el error sería cero; claro que la persona también sería cero: habría fenecido de una sobredosis de ítems; en matemáticas no, pero en psicología casi todo es cero en el límite.

Según los supuestos del modelo, si se tiene un test X y se aumenta su longitud n veces a base de ítems paralelos a los originales, la fiabilidad del nuevo test alargado viene dada por la conocida fórmula de Spearman-Brown, denominada con frecuencia «profecía de Spearman-Brown», dado que predice, profetiza, lo que le va a ocurrir a la fiabilidad del test al aumentar su longitud:

$$\rho_{XX'} = \frac{n\rho_{xx'}}{1 + (n - 1)\rho_{xx'}} \qquad [2.20]$$

donde:

$\rho_{XX'}$: Fiabilidad del test alargado.
$\rho_{xx'}$: Fiabilidad del test original.
n: Número de veces que se ha alargado el test.

Un caso particular de esta fórmula especialmente popular es cuando se duplica la longitud del test original, utilizado, por ejemplo, en el cálculo del coeficiente de fiabilidad por el método de las dos mitades. En dicho caso particular $n = 2$, quedando [2.20] reducida a:

$$\rho_{XX'} = \frac{2\rho_{XX'}}{1 + \rho_{XX'}} \qquad [2.21]$$

© Ediciones Pirámide

Ejemplo

Un test que consta de 20 ítems se aplicó a una muestra de personas, obteniéndose un coeficiente de fiabilidad de 0,60. ¿Qué fiabilidad tendría el test si se le añadiesen 15 ítems paralelos a los que ya poseía?

$$n = \frac{20 + 15}{20} = 1{,}75$$

$$\rho_{XX'} = \frac{(1{,}75)(0{,}60)}{1 + (1{,}75 - 1)(0{,}60)} = 0{,}724$$

La nueva fiabilidad del test alargado sería 0,724, frente al 0,60 del test original.

Todo lo dicho respecto a alargar el test puede aplicarse a acortar, en cuyo caso n será menor que 1.

Ejemplo

Un test que consta de 150 ítems tiene un coeficiente de fiabilidad de 0,90. Dado que resulta tediosamente largo, se le han suprimido 60 ítems. ¿Cuál será la fiabilidad del nuevo test acortado?

$$n = \frac{150 - 60}{150} = 0{,}60$$

$$\rho_{XX'} = \frac{(0{,}60)(0{,}90)}{1 + (0{,}60 - 1)(0{,}90)} = 0{,}844$$

Al suprimir los 60 ítems, la fiabilidad baja de 0,90 a 0,844.

De la fórmula general [2.20], se puede despejar n, lo que permite estimar cuánto habría que alargar (o acortar) un test para obtener una fiabilidad determinada:

$$n = \frac{\rho_{XX'}(1 - \rho_{xx'})}{\rho_{xx'}(1 - \rho_{XX'})} \qquad [2.22]$$

Ejemplo

La fiabilidad de un test de 40 ítems resultó ser 0,80. ¿Cuántos ítems habría que añadirle para que su fiabilidad fuese 0,90?

$$n = \frac{0{,}90(1 - 0{,}80)}{0{,}80(1 - 0{,}90)} = 2{,}25$$

Hay que alargar el test 2,25 veces; luego el nuevo test alargado tendrá (2,25)(40) = 90 ítems; habría que añadirle 50 ítems (90 – 40 = 50).

Obtención de la fórmula de Spearman-Brown

Como se ha podido comprobar a través de los ejemplos anteriores, el uso de la fórmula de Spearman-Brown es ciertamente sencillo. Un par de preguntas que surgen inmediatamente es: ¿funciona la fórmula?, ¿son correctas las «profecías» hechas a partir de ella? Nótese que los resultados proporcionados por esta fórmula son susceptibles de someterse a comprobación empírica, puede estimarse la fiabilidad mediante ella y luego calcular la fiabilidad empíricamente, contrastando los resultados. La abundante literatura al respecto parece indicar que la fórmula funciona bastante bien en general, siempre y cuando, claro está, los ítems añadidos sean paralelos a los previamente existentes.

Para la obtención de la fórmula de Spearman-Brown se parte del concepto general de coeficiente de fiabilidad proporcionado por el modelo lineal clásico, es decir, el cociente entre la varianza de las puntuaciones verdaderas y la de las empíricas. Ahora bien, si se alarga n veces el test, tanto la varianza de las verdaderas del nuevo test alargado como la de las empíricas se verán afectadas. En concreto (véase apéndice), la varianza verdadera original queda multiplicada por n^2, es decir, la varianza verdadera del test alargado n veces será: $n^2\sigma_v^2$. Por su parte, la varianza empírica del test alargado (apéndice) viene dada por

$$n\sigma_x^2[1 + (n - 1)\rho_{xx'}]$$

Por tanto, la fiabilidad del test alargado será el cociente entre ambas:

$$\rho_{XX'} = \frac{n^2\sigma_v^2}{n\sigma_x^2[1 + (n - 1)\rho_{xx'}]}$$

y simplificando

$$\rho_{XX'} = \frac{n\rho_{xx'}}{1 + (n - 1)\rho_{xx'}}$$

que es precisamente la fórmula propuesta.

© Ediciones Pirámide

Nótese que al aumentar n veces la longitud de un test su varianza verdadera aumenta proporcionalmente más que su varianza empírica, pues, mientras que la verdadera original resulta multiplicada por n^2, la empírica se multiplica por $n[1 + (n - 1)\rho_{xx'}]$, expresión cuyo valor solo se igualaría a n^2 en el caso de que $\rho_{xx'} = 1$, hecho poco probable en la práctica, por no decir imposible. Esta es precisamente la razón de que al aumentar la longitud aumente la fiabilidad, pues aumenta más el numerador (varianza verdadera) que el denominador (empírica).

Límite de $\rho_{xx'}$ cuando n tiende a infinito

Según Spearman-Brown:

$$\rho_{XX'} = \frac{n\rho_{xx'}}{1 + (n - 1)\rho_{xx'}}$$

bajando n al denominador

$$\rho_{XX'} = \frac{\rho_{xx'}}{[1 + (n - 1)\rho_{xx'}]/n}$$

y dividiendo cada sumando entre n y simplificando:

$$\rho_{XX'} = \frac{\rho_{xx'}}{(1/n) + (n - 1)\rho_{xx'}/n}$$

$$\rho_{XX'} = \frac{\rho_{xx'}}{(1/n) + \rho_{xx'} - (1/n)\rho_{xx'}}$$

Ahora bien, cuando n tiende a infinito: $1/n$ es cero; luego

$$\rho_{XX'} = \rho_{xx'}|\rho_{xx'} = 1$$

Es decir, la fiabilidad de un test tiende a uno a medida que se aumenta su longitud, alcanzando teóricamente ese valor para infinitos ítems. Se recomienda cierta precaución en el salto de las matemáticas a la psicología. Mejorar la fiabilidad de un test a base de aumentar su longitud puede ser útil y recomendable, por ejemplo, cuando el test original ya tiene una fiabilidad razonable con no muchos ítems, pero si el test ya tiene un número considerable de ítems y, sin embargo, es poco fiable, más que su longitud hay que ir pensando en cambiar los ítems, en construir otro.

6.3. Fiabilidad y nivel de las puntuaciones en el test

Hasta ahora se ha visto cómo se calculaba el coeficiente de fiabilidad y el error típico de medida para una muestra determinada, asumiendo implícitamente que sus valores eran comunes para todas las personas de la muestra, independientemente de sus puntuaciones en el test. Ahora bien, un test no siempre resulta igualmente preciso para todas las personas; su error típico de medida puede depender de la puntuación o nivel de las personas en el test. En este apartado se expone la forma de calcular el error típico de medida para distintos niveles de puntuaciones en el test. Pero antes de pasar a exponer el método de cálculo, veamos a qué se puede deber esta variación de los errores típicos de medida. La causa fundamental de que el error típico de medida no sea el mismo para cualquier nivel de puntuaciones radica en el tipo de ítems que componen el test. Por ejemplo, si ocurriese que la mayoría de los ítems fuesen de dificultad media, el test mediría con mayor precisión a las personas de nivel medio, es decir, los errores de medida tenderían a ser mayores para el caso de personas de alta y baja competencias en la variable medida. Por el contrario, si los ítems son en su mayoría de dificultad elevada, la prueba tenderá a dar mediciones más precisas para las personas de alto nivel, en detrimento de aquellas con puntuaciones medias o bajas. Casos extremos serían aquellos en los que todos los ítems fuesen tan difíciles, o tan fáciles, que no fuesen contestados, respectivamente, por casi nadie o por la mayoría. La variación del error típico tiende a agudizarse a medida que aumenta la amplitud del rango de las puntuaciones en la variable medida. Además de la naturaleza de los ítems, pueden existir otros factores de carácter secundario que contribuyan también a que el error típico de medida no afecte por igual a todas las personas de la muestra, tales como unas instrucciones inadecuadas que induzcan a las personas con poco nivel a contestar al azar ante cualquier pregunta que

© Ediciones Pirámide

desconozcan, u otros por el estilo derivados de una aplicación incorrecta de la prueba.

Ante esta situación no parece apropiado usar el mismo error típico de medida para todas las personas, por lo que se recurre a la utilización de distintos errores típicos de medida en función de la cuantía de las puntuaciones de las personas en el test. Dado que no se puede generalizar de unos test a otros, ni de unas muestras a otras, habrá que calcular empíricamente en cada caso los errores típicos de medida correspondientes.

La forma más clásica de llevar a cabo el cálculo de los errores típicos de medida para los distintos niveles de puntuaciones fue propuesta por Thorndike (1951). Consiste en dividir las puntuaciones en varios niveles o categorías y calcular el error típico de medida para cada una de ellas. No se puede hablar de un número idóneo de categorías, que dependerá en gran medida del número de personas de la muestra. Un número mínimo podrían ser tres, puntuaciones bajas, medias y altas, pero si se dispone de suficientes sujetos, puede incrementarse el número de niveles, explorando de ese modo los errores con mayor exhaustividad a lo largo del rango de la variable medida. Si se dispusiese de las suficientes personas, podrían incluso hacerse tantas categorías como puntuaciones posibles en el test. Estrictamente hablando, los niveles deberían establecerse a partir de las puntuaciones verdaderas, pero en la práctica solo se dispone de las empíricas.

Si se dispone de una sola aplicación del test, una vez establecidas las categorías de las puntuaciones, se calcula el error típico de medida para cada una de ellas. Para ello se dividen las puntuaciones de cada persona en dos mitades (pares e impares, por ejemplo) y se calcula la desviación típica de las diferencias entre ambas. El resultado sería el error típico de medida para cada uno de los niveles de puntuaciones:

$$\sigma_e = \sigma_{(p-i)}$$

Nótese que si el test fuese perfectamente fiable, las puntuaciones de ambas mitades coincidirían, no habría errores de medida, el error típico de medida sería nulo y la fiabilidad perfecta. A medida que aumentan las diferencias entre las dos mitades, se incrementa el error típico de medida. Esta forma de conceptualizar el error típico de medida fue utilizado por Rulon (1939) para obtener su fórmula del coeficiente de fiabilidad (véase la fórmula 2.29). Si se denomina e al error del test global y e_1 y e_2 a los errores de cada una de las mitades, bajo los supuestos del modelo clásico, es fácil demostrar que la varianza de los errores globales del test es igual a la suma de las varianzas de los errores de cada una de las dos mitades:

$$\sigma_e^2 = \sigma^2(e_1 - e_1) = \sigma_{e1}^2 + \sigma_{e2}^2$$

Si se dispone de dos formas paralelas, o de dos aplicaciones del test, se forman las categorías a partir de la suma de las puntuaciones de cada persona en ambas formas y luego se procede al cálculo del error típico de medida para cada categoría. El error típico se obtiene calculando la desviación típica de las diferencias entre las dos formas del test y dividiendo el resultado por $\sqrt{2}$

$$\sigma_e = \frac{\sigma(X_1 - X_2)}{\sqrt{2}} \qquad [2.23]$$

En la fórmula 2.16 se expresa la desviación típica de las diferencias entre dos formas paralelas de un test; nótese cómo para obtener a partir de ella el error típico de medida hay que dividirla por $\sqrt{2}$.

A continuación se ilustran los cálculos anteriores mediante un ejemplo numérico.

Ejemplo

Se aplicó un test de aptitud espacial a una muestra de 12 personas. Las puntuaciones globales en el test, así como las obtenidas en las mitades par e impar, se ofrecen a continuación. Veamos cómo se calcula el error típico de medida para tres niveles de la variable medida: bajo, medio y alto.

	Personas	Global	Par	Impar	Diferencia ($P - I$)
Nivel bajo	A	6	4	2	2
	B	7	4	3	1
	C	9	5	4	1
	D	10	5	5	0

© Ediciones Pirámide