Avances en Medición, 1 (1), 17-36
2003

DESARROLLOS RECIENTES EN PSICOMETRÍA

Juana Gómez-Benito
Universitat de Barcelona, España

M. Dolores Hidalgo-Montesinos
Universidad de Murcia, España

Resumen

En casi un siglo de desarrollo de la Psicometría, desde que a comienzos del siglo XX Spearman esbozara las primeras ideas acerca de la teoría clásica de tests, el camino recorrido por la misma ha sido largo, y desde esos inicios hasta la actualidad se han introducido numerosos avances que han cambiado los enfoques sobre la medida en las ciencias sociales y educativas. También se han introducido cambios en la forma de entender los conceptos principales de la medida, tales como fiabilidad, validez, sesgo del test y del ítem, invarianza de parámetros, dimensionalidad y objetividad de la medida. El objetivo del presente trabajo es conducir al lector hacia los desarrollos más recientes en Psicometría, introduciendo una de las teorías psicométricas de mayor impacto científico, la IRT, y presentando algunos de los temas aplicados de mayor actualidad.

Palabras clave: Teoría de la Respuesta al Item, Teoría de la Generalizabilidad, Teoría Clásica de Tests, Funcionamiento Diferencial del Item, Traducción de Tests.

Abstract

Psychometry has come a long way in the hundred years or so since, at the beginning of the twentieth century, Spearman outlined the first ideas concerning classical test theory. Since then, numerous advances have been made that have changed the way in which the social and educational sciences have approached measurement. There have also been changes in the way in which the main measurement concepts –such as reliability, validity, test and ítem bias, parameter invariance, the dimensionality and objectivity of measurement– are understood. The present paper considers the most recent developments in psychometry and introduces one of the most scientifically important psychometric theories, IRT. Some of the latest applied topics are also discussed.

Key words: Item Response Theory, Generalizability Theory, Classical Test Theory, Differential Item Functioning, Tests Translation.

Desde que a comienzos del siglo XX Spearman delineara los primeros trazos sobre lo que se llamaría teoría clásica de los tests (TCT), la Psicometría ha ido incorporando distintos enfoques sobre la medida que han introducido cambios paulatinos en los principales conceptos de la medición, como fiabilidad, validez, sesgo del test y del ítem, invarianza de parámetros, dimensionalidad y objetividad de la medida. El último tramo de esta evolución ha estado marcado por el desarrollo de nuevos modelos de medida, principalmente en el marco general

de lo que se conoce como *Teoría de Respuesta al Ítem* (Item Response Theory, IRT) (Hambleton y Swaminathan, 1985; Lord, 1980; Van der Linden y Hambleton, 1997) y en menor medida en el ámbito de la *Teoría de la Generalizabilidad* (Generalizability Theory, GT) (Brennan, 1983; Cronbach, Gleser, Nanda y Rajaratnam, 1972; Shavelson y Webb, 1991).

El uso y aplicación de ambas teorías, IRT y GT, ha modificado algunas de las reglas tradicionales en la medida invitando a los constructores de tests y a los que se dedican a la evaluación -psicólogos o educadores- a que se formen en estos nuevos modelos de medida. En este sentido, los trabajos de Embretson (1996) y Embretson y Reise (2000) son una primera referencia para conocer qué viejas reglas han sido sustituidas por otras nuevas. Si de la TCT se deriva que tests más largos son más fiables que tests cortos, con la IRT tests cortos pueden ser más fiables que tests más largos. Si con la TCT, en la elaboración y selección de ítems para un test, las características estimulares de los mismos no son importantes y se da más relevancia a las propiedades del conjunto de ítems (test) que a la contribución individual de cada ítem, con la IRT las características estimulares de los ítems pueden relacionarse con las propiedades psicométricas del test. Otras viejas normas de medida también han cambiado, tales como: 1) para conseguir estimaciones insesgadas de las propiedades de los ítems se necesitan muestras representativas, 2) el error típico de medida es el mismo para todas las puntuaciones en una población de sujetos, 3) no es factible combinar, en un mismo test, ítems con distintos formatos sobre todo porque pueden ocasionar un impacto no balanceado en las puntuaciones del test y 4) las puntuaciones de los tests adquieren su significado, se interpretan, por comparar su posición respecto a un grupo normativo. Ahora, con los nuevos modelos: 1) es posible obtener estimaciones insesgadas de las propiedades de los ítems trabajando a partir de muestras no representativas de la población, 2) el error de medida se puede estimar condicionado a cada una de las puntuaciones (niveles del rasgo) del test y se puede generalizar a través de las poblaciones, 3) es posible combinar en un mismo test distintos formatos de ítems y obtener puntuaciones adecuadas para los sujetos, y 4) las puntuaciones de los tests adquieren significado no por su posición a lo largo del continuo de habilidad, sino por su posición respecto al continuo de dificultad del ítem, es decir, respecto a su distancia de los ítems.

Vistas así las cosas, el objetivo del presente trabajo es exponer las pautas de esta evolución hacia los desarrollos más recientes en Psicometría, haciendo hincapié en la teoría de mayor impacto psicométrico en estos momentos, la TRI, y presentando algunos de los temas aplicados de mayor actualidad.

Lo clásico en Psicometría

La TCT ha sido durante gran parte del siglo XX, el modelo estrella en el que se ha fundamentado la medida psicológica y educativa a través de tests. Los trabajos pioneros de Spearman (1907) a principios del siglo XX establecieron las bases de esta teoría que se

desarrolla en el fermento de tres ingredientes (Traub, 1997): a) reconocimiento de la presencia de errores en la medida, b) la concepción de que el error de medida es una variable aleatoria, y c) la idea de correlación entre variables y de un coeficiente para expresarla, junto al desarrollo de la distribución normal bivariada. Así, los supuestos de la TCT se articulan en torno a las propiedades de la distribución de los errores de medida y su independencia de las puntuaciones verdaderas. La finalidad es obtener una cuantificación del error asociado a un proceso de medida, o por el contrario una estimación de la precisión del instrumento. La TCT utiliza el concepto de *tests paralelos* para desarrollar el de precisión de la medida o *coeficiente de fiabilidad*. De este modo, el coeficiente de fiabilidad de un test vendría dado por la correlación entre las puntuaciones obtenidas en dos tests paralelos.

Durante la primera mitad del siglo XX los esfuerzos de la Psicometría se dedicaron a establecer los factores que afectan al coeficiente de fiabilidad de un test, así como los distintos procedimientos que permiten estimar el coeficiente de fiabilidad. En la década de los 50 surge la primera formalización de la TCT en el libro de Gulliksen (1950) y a finales de la década de los 60 aparece el libro de Lord y Novick (1968) que marca un hito tanto para la fundamentación estadística de la TCT como en el desarrollo de nuevos modelos de medida, en concreto, para el desarrollo y aplicación de la IRT.

A pesar de su simplicidad y fácil adaptación a diversos campos de la Psicología, la TCT ha recibido críticas referidas tanto a los conceptos que maneja como a los supuestos en los que se fundamenta. En la tabla 1 aparecen resumidas las críticas más importantes a esta teoría.

Tabla1.
Limitaciones más importantes de la TCT.

1. La definición de tests paralelos es muy restrictiva.
2. Homoscedasticidad de los errores de medida.
3. Dependencia muestral de las propiedades psicométricas de los tests.
4. Dependencia muestral de las características de los ítems.
5. Dependencia de X (puntuación observada) de la longitud del test.

Una práctica habitual en el trabajo con tests psicológicos y/o educativos, ya sea con fines aplicados como de investigación, resulta en la comparación de las puntuaciones de los sujetos obtenidas en diferentes formas de un test. En estos casos es necesario algún tipo de igualación de sus puntuaciones antes de llevar a cabo cualquier comparación. Desde la TCT una primera solución al problema requiere utilizar el concepto de paralelismo de la medida. Sin embargo, la definición de tests paralelos es demasiado restrictiva, tanto que es muy difícil construir empíricamente tests que cumplan este supuesto, es decir, que cumplan las condiciones de igualdad de las medias, igualdad de las varianzas empíricas e igualdad de los coeficientes de fiabilidad en las distintas formas del test. Esto supone una seria limitación del modelo clásico,

así, desde la psicometría se ha dedicado un gran esfuerzo al desarrollo de procedimientos para la igualación de tests, siendo en la década de los 80 el libro editado por Holland y Rubin (1982) un buen referente de los problemas asociados con la comparación de puntuaciones y de los métodos clásicos de equiparación junto con los basados en la IRT. La importancia del tema es tal que a mediados de la década de los noventa Kolen y Brennan (1995) publican un manual que recoge una revisión muy actualizada del tema (ver también en castellano Navas, 1996 y 2000). Pero, aunque los procedimientos de igualación permiten la comparación de puntuaciones obtenidas de tests, no necesariamente paralelos, un problema inherente a todo proceso de igualación es el error asociado al mismo y la obtención de una estimación del mismo (Zeng, 1993; Zeng y Cope, 1995). Tal y como apuntan Peterson, Marco y Stewart (1982), el error de igualación se deja influir por diferencias en las propiedades psicométricas de los tests a igualar, siendo más elevado cuando se trabaja con tests que difieren en nivel de dificultad. Esta temática es una de las que más interés ha generado en la última década, junto a los trabajos que tratan de responder a la cuestión de ¿qué método de equiparación es mejor en cada situación de medida?. Sin embargo, otras líneas de investigación han aparecido en los últimos años, como respuesta a los nuevos formatos de los instrumentos de medida con los que se trabaja: el desarrollo de procedimientos de equiparación para tests multidimensionales, para tests construidos bajo la perspectiva de lo que se denomina *medición auténtica* (*performance assessment*) o para tests con formatos de respuesta politómica marcan la pauta del trabajo actual en equiparación de puntuaciones.

La TCT considera que el error típico de medida es constante (homoscedástico) para las puntuaciones obtenidas en una misma población de sujetos. Sin embargo, en la práctica las puntuaciones de error pueden estar sesgadas en los extremos de la distribución, es decir, para sujetos con puntuaciones verdaderas muy altas o muy bajas. Dado que el error típico de medida varía en función de la habilidad (Feldt, Steffen y Gupta, 1985), es necesario considerar estas variaciones individuales cuando se trabaja a un nivel de análisis individual, o lo que es lo mismo, cuando se interpretan las puntuaciones de los sujetos. Aunque desde la propia TCT se han propuesto varias aproximaciones para estimar el error típico de la persona (Feldt y Brennan, 1989; Woodruff, 1990), la IRT (Hambleton y Swaminathan, 1985; Lord, 1980) ofrece una solución más eficaz a este problema.

Otros puntos críticos en la TCT se refieren a la dependencia muestral de las características psicométricas tanto del test como de los ítems. En la TCT las propiedades del test (coeficiente de fiabilidad, error típico de medida, etc.) no son invariantes a través de las submuestras de sujetos en las que el test se administra, dependen de la variabilidad del grupo, siendo esto también aplicable a las características de los ítems, tales como su índice de dificultad y su índice de discriminación(correlación ítem-test). En función de ello, el coeficiente de fiabilidad será mayor cuando más heterogéneo sea el grupo; además la fiabilidad también depende del tamaño del test, a mayor número de ítems en un test mayor precisión del mismo (fórmula de la profecía de Spearman-Brown). Por otro lado, las características de los ítems dependen de la

habilidad promedio del grupo donde se obtienen y de su variabilidad, así el índice de dificultad estimado para un ítem será más elevado en grupo de habilidad promedio alta que en un grupo con un nivel de habilidad media baja. La invarianza de la medida es un requisito imprescindible asociado a cualquier instrumento de medida (Engelhard, 1992) e implica que las propiedades de los ítems no varían a través de los subgrupos de sujetos a los que se les administre el test, es decir, no son función de características grupales tales como habilidad, género, raza o clase social. La invarianza también supone que las propiedades de los ítems no están afectadas por la inclusión o exclusión de otros ítems en el test. El cumplimiento de esta propiedad permite que los sujetos sean comparados independientemente de la muestra de ítems que se haya utilizado para evaluarlos.

Por último, la puntuación observada (X) de un sujeto en un test depende del tamaño del test. En el modelo clásico, X se obtiene como la suma de las puntuaciones observadas en cada uno de los ítems del test, sin tener en cuenta las características diferenciales de los sujetos en el rasgo, es decir, en la TCT se asume que los ítems de un test son paralelos (miden por igual el rasgo psicológico) y la puntuación del sujeto en el test se obtiene como una combinación lineal de los ítems. Este sistema de puntuación aunque sencillo de implementar es poco realista desde un punto de vista práctico.

Estas críticas han impulsado la investigación de otros modelos de medida como respuesta a los problemas de aplicación de la TCT, siendo la TG y la IRT las alternativas más importantes con las que hoy en día se trabaja.

Las características esenciales de la TG se pueden resumir en los siguientes puntos (Brennan, 1997; Shavelson y Webb, 1991): 1) diferencia entre múltiples fuentes de error asociadas a todo proceso de medida -en contraposición a la TCT que considera el error de medida de forma compacta y unitaria-, estimando esas fuentes de error a través del ANOVA y proporcionando un mecanismo para optimizar la fiabilidad de la medida, 2) sustituye el concepto de puntuación verdadera por el de puntuación universo, 3) utiliza el término *facetas* para referirse a una fuente de variación de error (ítems, jueces, ocasiones de medida, etc.) y el de *condiciones* para referirse a los niveles de la faceta medida, 4) relaja el supuesto de ítems o tests paralelos por el de ítems o tests aleatoriamente paralelos, 5) reemplaza el concepto de fiabilidad de la TCT por el de *generalizabilidad*, 6) distingue dos tipos de varianza de error: absoluta y relativa, 7) diferencia entre estudios G (generalización) y D (decisión), los primeros estiman la magnitud de las fuentes potenciales de error de medida y los segundos usan la información de los primeros para obtener medidas que minimicen el error, explorando el efecto en la precisión de la medida bajo distintas condiciones o situaciones preestablecidas.

La GT no está libre de críticas (ver Rozeboom, 1978), siendo una de las limitaciones más importantes de la misma la dependencia de las puntuaciones universo de los ítems que se han empleado. La difusión de la GT en ámbitos aplicados de la medida psicológica estuvo mermada por otro fuerte competidor, la IRT, pero en los últimos veinte años se puede

comprobar en la literatura cómo ha ido ganando adeptos. Los trabajos de Gillmore (1983) en la evaluación de programas; Anguera (1990), Blanco (1989, 2001) y Marcoulides (1989) en la investigación observacional, Brennan (1992, 2000), Brennan, Gao y Colton (1995), Cronbach, Linn, Brennan y Haerteel (1997), Fitzpatrick, Guemin y Furong (2001), en la evaluación educativa y medida auténtica, por poner algunos ejemplos, muestran el potencial de estos modelos; recientemente también se viene aplicando al ámbito de los recursos humanos y la evaluación psicológica.

Teoría de Respuesta al Ítem: su contribución a la Psicometría

La investigación sobre la medición de variables psicológicas ha estado dirigida siempre a mejorar la evaluación mediante tests y a solucionar los problemas encontrados con su uso. En los últimos 25 años, la IRT se ha vislumbrado como una alternativa que ofrece soluciones viables a algunas de las limitaciones teóricas y prácticas de la TCT. Además, la IRT no es sólo la teoría psicométrica subyacente a muchos tests de hoy en día, sino que ofrece múltiples y relevantes aplicaciones de investigación. Todo esto ha hecho de esta teoría una de las más prometedoras en la fundamentación de la medida psicológica y educativa. Un vistazo a las bases de datos más importantes en Psicología y Educación (PsycLIT y ERIC) así como en el ámbito de la Medicina (MEDLINE) permite constatar cómo el número de trabajos (artículos, libros, tesis y comunicaciones a congresos) que incluyen el término Teoría de Respuesta al Ítem es con diferencia mayor que el de trabajos que se refieren a la Teoría Clásica de Tests y a las aplicaciones de la Teoría Generalizabilidad. Además, en los últimos años se han desarrollado un buen número de trabajos de divulgación de la IRT en ámbitos profesionales en los que se venía aplicando la TCT (Cella y Chang, 2000; Embretson y Prenovost, 1999; Embretson y Reise, 2000; Harvey y Hammer, 1999; Hays, Morales y Reise, 2000; Yen y Edwardson, 1999).

Perspectiva histórica

Aunque la IRT se puede considerar un modelo de medida moderno, su desarrollo comienza a trazarse en el primer cuarto del siglo XX con los trabajos de Thurstone que intentan sintetizar la psicofísica y la teoría de tests. Los antecedentes de la IRT también pueden ser trazados a partir de los trabajos de Fisher en el campo del bioensayo y la toxicología. Sin embargo, las referencias más directas se encuentran en los trabajos a finales de los años treinta y principios de los cuarenta de Richardson (1936), Lawley (1943), Guttman (1944) y Tucker (1946) en los que se desarrolla el concepto de Curva Característica del Ítem (Item Characteristic Curve, ICC). A mediados de siglo, Lazarsfeld (1950) introduce, en el campo de las actitudes, el principio de independencia local de ítems y el concepto de clase latente, proponiendo un modelo de distancia latente y un modelo lineal, y Lord (1952) describe el modelo de ogiva normal de dos parámetros propiciando las primeras aplicaciones de estos modelos. Durante estos años Rasch desarrolla dos modelos basados en la distribución

de Poisson para tests de lectura y un modelo basado en la función de logística para tests de inteligencia y rendimiento. Este último modelo, denominado por Rasch *Structural Model for Items in a Test*, se conoce actualmente como Modelo de Rasch (Rasch, 1960; Van der Linden y Hambleton, 1997). Durante los años cincuenta y de forma independiente a Rasch, Birnbaum trabaja también en modelos de respuesta al ítem a partir de la función logística. En la década de los sesenta aparece el libro de Rasch y más tarde el trabajo de Birnbaum (1968) sobre los modelos logísticos de 2-p y 3-p, como capítulos en el libro de Lord y Novick (1968). A partir de los años setenta el interés se centra en el desarrollo y búsqueda de métodos de estimación de parámetros (Bock y Lieberman, 1970; Wood, Wingersky y Lord, 1976), y su implementación en ordenador, por lo que aparecen los primeros programas de estimación de parámetros como BICAL (Wright y Mead, 1976) y LOGIST (Wood, Wingersky y Lord, 1976). También se empiezan a atisbar las posibilidades prácticas de estos modelos, como queda recogido en el número especial de la revista *Journal of Educational Measurement* (1977, vol. 14, nº 2). En la década de los 80, coincidiendo con el desarrollo de los ordenadores, el interés sobre la IRT es máximo como así lo pone de manifiesto el libro de Lord (1980), donde se tratan las aplicaciones de la IRT, y los números especiales de las revistas *Applied Psychological Measurement* (1982, vol. 9, nº 4), *Applied Measurement in Education* (1988, vol. 1, nº 4) e *International Journal of Educational Research* (1989, vol. 13, nº 2). Aparecen distintos manuales orientados principalmente a los modelos de teoría de respuesta al ítem para ítems dicotómicos y al modelo de Rasch (Andrich, 1988; Baker, 1985; Hambleton y Swaminathan, 1985; Hulin, Drasgow y Parsons, 1983; Weiss, 1983; Wright y Stone, 1979). El número de trabajos sobre IRT y aplicaciones aumenta considerablemente, encontrándose hacia finales de los 80 y principios de los 90 diversas revisiones sobre el tema (Baker, 1987, 1992; Goldstein y Wood, 1989; Hambleton, 1990; Jones y Appelbaum, 1989; Muñiz y Hambleton, 1992; Traub y Lam, 1985). En la última década del siglo XX y principios del siglo XXI la producción científica con relación a la IRT no ha decrecido y el interés en este tipo de modelos sigue aumentando. El libro de Fischer y Molenaar (1995), exclusivamente referido al modelo de Rasch y sus derivaciones, muestra el potencial de este modelo para analizar situaciones de medida más allá de las típicas que se producen como resultado de la administración de un test. El gran interés en los modelos de Rasch y sus aplicaciones (Rasch Measurement) queda recogido en las reuniones que cada dos años se celebran sobre este tema en Estados Unidos (International Objective Measurement Workshop) cuyos trabajos han visto la luz en los libros editados por Wilson (1992, 1994), Engelhard y Wilson (1996), Wilson, Engelhard y Draney (1997) y Wilson y Engelhard (en prensa). Por su parte, el libro de Van der Linden y Hambleton (1997) y la revisión de Barbero (1999) muestran el extenso abanico de modelos de IRT, mientras Bock (1997) ofrece una excelente revisión histórica sobre el desarrollo de la IRT. En castellano cabe citar los manuales de López-Pina (1995) y Muñiz (1990, 1997). Por último, ya en los albores del nuevo siglo, el libro de Embretson y Reise (2000) supone una apuesta por el uso y difusión de la IRT en el ámbito de la evaluación psicológica.

Características y aportaciones

¿Cuáles son las ventajas fundamentales de la IRT que han hecho de la misma uno de los modelos de medida más usados en los últimos años?. En primer lugar, el ajuste de un modelo de IRT proporciona estimaciones invariantes de las propiedades psicométricas de los ítems, así como de las características psicológicas de los sujetos, es decir, que los parámetros que caracterizan al ítem no dependen de la muestra particular de sujetos utilizada y que los parámetros que caracterizan al sujeto no dependen de la muestra particular de ítems utilizada (Hambleton y Swaminathan, 1985; Lord, 1980). En segundo lugar, la IRT proporciona una descripción detallada del comportamiento individual de los ítems a lo largo del continuo de habilidad, cuya representación gráfica se denomina ICC o Función de Respuesta al Ítem (Item response function, IRF). Otra de las características ventajosas de la IRT es que permite obtener índices de precisión no sólo del test sino también del ítem condicionados al nivel de habilidad del sujeto, es decir, el ajuste de un modelo de IRT a un test facilita la estimación de la precisión (o por el contrario, del error de medida) en cada nivel del rasgo, donde se admite la variación de la misma a lo largo del continuo de habilidad. El error típico de medida, contrario al proporcionado en la TCT, se presenta condicionado al nivel de habilidad específico del sujeto, siendo básico para describir la calidad de un test y también de cara a la interpretación individual de las puntuaciones obtenidas por los sujetos. Sin embargo, pueden encontrarse situaciones donde los errores típicos de la TCT y de la IRT serán similares, cuando el nivel de dificultad del test coincida con el nivel medio de habilidad del grupo y además la distribución de habilidad para esa población cumpla el supuesto de normalidad. En cualquier caso, la IRT elabora los conceptos de Función de Información del Ítem (Item information function) y Función de Información del Test (Test information function) que serán las herramientas de trabajo para la selección de ítems y construcción de tests en el contexto de este modelo de medida.

En el marco de la IRT se proponen un conjunto de índices para valorar el ajuste de personas (person-fit) que evalúan la calidad del modelo de IRT ajustado con respecto a los individuos así como el significado de las interpretaciones acerca de las puntuaciones obtenidas. En general, estos índices comparan el patrón de respuestas del sujeto con respecto al patrón de respuestas ideal que se obtendría del ajuste del modelo, valorando la consistencia o inconsistencia del mismo, con la finalidad de identificar personas cuyos patrones de respuestas en el test sean improbables. El desarrollo de los índices de ajuste de personas se llevó a cabo en el contexto de la medición educativa (Levine y Rubin, 1979; Meijer y Sitjsma, 1995) para detectar sujetos que han copiado, que responden al test prestando poca atención o de forma aleatoria, o que por ser extremadamente creativos reinterpretan los ítems fáciles y los fallan; también para detectar errores al pasar los resultados del test a la hoja de respuestas, o problemas en la propia construcción del test. En la actualidad se trabaja con estos índices también en el ámbito de la medida de variables psicológicas y de personalidad, con la finalidad de evaluar patrones de respuesta que no se comportan en el mismo sentido que lo

esperado según el rasgo evaluado (Reise y Waller, 1993). En este contexto algunas de las posibles fuentes de desajuste serían deseabilidad social, desmotivación al responder al test, falsificación de las respuestas, aquiescencia, o la variación en el rasgo (en el sentido de si la expresión del rasgo de un sujeto es comparable a otras expresiones del mismo rasgo desde un punto de vista cualitativo). En general, este ámbito de investigación ocupa tanto a teóricos como prácticos, puesto que con estos índices es posible detectar fenómenos de interés que tienen que ver con el proceso de responder a un test (Meijer y Nering, 1997).

Entre las características principales de la IRT se encuentra precisamente que es una aproximación a la medida basada en modelos y por lo tanto dispone de técnicas para evaluar el ajuste del modelo. Es más, las ventajas asociadas con el uso de un modelo de IRT (independencia muestral, banco de ítems, etc.) sólo se pueden obtener cuando se comprueba que el ajuste entre el modelo y los datos es satisfactorio. La evaluación del ajuste del modelo se realiza a tres niveles: a) evaluando el ajuste entre los valores observados y los esperados bajo el modelo estimado, usando pruebas de bondad de ajuste referidas a los ítems (basadas en ji-cuadrado, en el análisis de residuales y en el estadístico de razón de verosimilitud) y pruebas de bondad de ajuste del modelo basadas en la comparación entre modelos, b) evaluando si los datos observados satisfacen los supuestos del modelo que interesa ajustar (dimensionalidad, independencia local, igualdad de los parámetros de discriminación, ausencia de adivinación y ausencia de velocidad) y c) evaluando si se cumplen las ventajas derivadas del modelo. Revisiones sobre el tema se pueden encontrar en Hambleton y Swaminathan (1985), López Pina e Hidalgo (1996) y Traub y Lam (1985).

Por otro lado, el abanico actual de modelos de IRT que se encuentran en la literatura es bastante amplio y se caracterizan por ser más o menos restrictivos en las especificaciones de partida. Así, los modelos difieren en si la respuesta al ítem viene explicada por uno o más rasgos latentes (modelos unidimensionales vs. multidimensionales), la CCI está definida o no teóricamente (modelos paramétricos vs. no-paramétricos), la CCI está definida por uno o más parámetros (modelos de Rasch y derivaciones vs. modelos multiparamétricos), formato de la respuesta (dicotómicos, politómicos, continuo, tiempo de respuesta), etc. Aunque pueden encontrarse varias propuestas de clasificación de estos modelos, el trabajo de Mellenbergh (1994) proporciona un marco integrado de presentación de los mismos.

La IRT supuso un gran paso en la elaboración de tests adaptativos que combinados con el desarrollo de los ordenadores dio lugar a los tests adaptativos informatizados (Van der Linden y Zwarts, 1989). Además, la IRT proporciona un marco integrado para analizar el Funcionamiento Diferencial del Ítem (Differential Item Functioning, DIF) y el Funcionamiento Diferencial del Test (Differential Test Functioning, DTF); un ejemplo de esto se puede ver en los trabajos de Raju, Van der Linden y Fleer (1995), Fleer, Raju y Van der Linden (1995), Oshima, Raju, Flowers y Monaco (1995) y Flowers, Oshima y Raju (1996, 1997). En la tabla 2 aparecen resumidas las aportaciones más relevantes de la IRT.

Tabla 2
Aportaciones más relevantes de la IRT

1. Invarianza de los parámetros de los ítems
2. Invarianza de los parámetros de habilidad de los sujetos
3. Curva Característica del Ítem
4. Error de medida condicionado al nivel de habilidad
5. Función de información del ítem y del test
6. Medidas de evaluación de la calidad del perfil de respuesta (ajuste de la persona)
7. Medidas de evaluación del ajuste individual de los ítems al modelo
8. Amplia gama de modelos que responden a diferentes situaciones de medida
9. Facilita el desarrollo de tests adaptativos computerizados
10. Facilita la evaluación del DIF y del DTF

Algunos temas candentes en la medida psicológica y educativa

Aunque actualmente son numerosos los temas pendientes en Psicometría que están generando investigación (fundamentación de la medida, estimación de parámetros, evaluación del ajuste de modelos, estimación de modelos en IRT, entre otros), el presente trabajo se centra en tres de ellos. En primer lugar, el sesgo de los tests psicológicos, dado que afecta a su validez y a la toma de decisiones con los mismos, en segundo lugar, la adaptación de cuestionarios de unas culturas a otras y, en tercer lugar, los tests adaptativos informatizados cuyo uso se está consolidando.

Evaluación y detección del funcionamiento diferencial del test y del ítem

Uno de los temas que más investigación e interés político y social ha suscitado en la última mitad del siglo XX, es el de la detección del sesgo del ítem y/o del test. Que un test/ítem está sesgado implica que el test o el ítem favorece injustamente a un grupo sobre otro, es decir, se encuentra disparidad en el comportamiento del mismo en función de alguna variable grupal (p.e., raza, sexo, nivel socioeconómico, bagaje cultural y dominio lingüístico) que no resulta relevante para el constructo que trata de medir el test. El término más técnico para referirse a este problema es Funcionamiento Diferencial del Item: un ítem presenta DIF cuando sus propiedades estadísticas varían en función de las características del grupo que lo ha contestado, siempre que los grupos comparados estén equiparados en el rasgo que mide el test. Normalmente, los estudios de DIF se basan en la comparación de dos grupos de individuos, a los que se denomina grupo focal y grupo de referencia. El primero de ellos, la mayoría de las veces minoritario, resulta generalmente perjudicado por la existencia de DIF en el test,

mientras que el segundo se toma como un estándar con respecto al cual se compara el grupo focal. Es importante no confundir DIF con Impacto; este último atañe a diferencias reales entre grupos en el constructo medido por el test. Por el contrario, el DIF hace referencia a diferencias en los estadísticos del ítem después de que los sujetos hayan sido igualados en el rasgo que mide el test, lo que equivale a una diferencia no esperada entre grupos que se supone tienen la misma habilidad. Así, cuando se detecta impacto entre grupos no es lícito asumir directamente que el test está sesgado y favorece a uno de los grupos sobre otro; esto último sólo es posible constatarlo llevando a cabo un estudio de DIF.

En la literatura psicométrica se encuentran una amplia gama de procedimientos para detectar DIF y DTF (ver revisiones teóricas en Fidalgo, 1996; Gómez e Hidalgo, 1997; Herrera, Sánchez y Gómez, 2001; Hidalgo y Gómez, 1999; Hidalgo y López, 2000; Millsap y Everson, 1993; Potenza y Dorans, 1995). En términos generales estos procedimientos se pueden dividir en dos amplias categorías: 1) los que utilizan como criterio de equiparación de los grupos la puntuación observada en el test (estadístico Mantel-Haenszel, procedimiento estandarizado, modelos de regresión logística, modelos loglineales, análisis discriminante logístico, método delta-plot) y los que utilizan la habilidad latente estimada bajo algún modelo de IRT (estadístico de Lord, medidas de área, comparación de modelos, SIBTEST). Probar la efectividad de estas técnicas bajo distintas condiciones de DIF (porcentaje de ítems con DIF en el test, cantidad DIF, tipo de DIF, presencia o no de impacto entre grupos, tipo de DIF, tamaño muestral de los grupos bajo estudio, distintos formatos de respuesta de los ítems, presencia de multidimensionalidad) ha sido y es una de las tareas actuales de los psicómetras, cuyo objetivo principal es conocer la potencia y tasa de error tipo I de los procedimientos que se dispone, con la finalidad de proporcionar al práctico luz en la selección de un procedimiento para detectar DIF. Un resumen de las ventajas e inconvenientes de estas técnicas aparece en Gómez e Hidalgo (1997) y en Hidalgo y Gómez (1999).

Si bien se dispone de una amplia literatura acerca de la efectividad de las distintas técnicas para detectar DIF, la cuestión de la interpretación del DIF, es decir, la respuesta a la pregunta ¿cuáles son las causas por las que un ítem o un test funciona de forma diferente para dos o más grupos de una población?, está todavía pendiente. Algunos logros ya se han conseguido en esta dirección, y una de las explicaciones a este problema sería la multidimensionalidad de los ítems, es decir, que los ítems que presentan DIF miden alguna dimensión latente (habilidad ruido) que no es la que se pretende medir con dicho test. Sin embargo, avanzar en el estudio de las causas del DIF es todavía la asignatura pendiente en este ámbito de investigación, y enlaza directamente con la problemática de la validez del test.

Traducción y adaptación de cuestionarios

La adaptación y traducción de tests de una lengua y/o cultura a otra es una práctica bastante habitual (Oakland y Hu, 1992), y una revisión de las revistas en materia psicológica nos pone de manifiesto los numerosos equipos de investigación en las diversas temáticas de la

psicología (evolutiva, clínica, social,...) que dedican esfuerzo a la adaptación de instrumentos de una lengua a otra. La Comisión Internacional de Tests (International Test Commission, ITC), que viene trabajando en esta temática en los últimos años, apunta que sólo durante el año 1992 algunos tests desarrollados en USA fueron traducidos y adaptados a más de 50 lenguas (Oakland, Poortinga, Schlegel y Hambleton, 2001). Además, hay un gran interés en la realización de estudios transculturales que permitan la comparación de los resultados obtenidos en distintas culturas y el establecimiento de pautas comunes y diferentes entre culturas.

El proceso de traducción y adaptación de un test requiere algo más que la traducción del test de la lengua origen a la lengua destino, es necesario asegurar que las puntuaciones obtenidas con el test traducido son equivalentes a las obtenidas con el test original. Para alcanzar esa equivalencia hay que considerar cuatro aspectos del proceso (Hambleton, 1994a): 1) el contexto cultural donde se va a realizar la adaptación, 2) aspectos técnicos del propio desarrollo y adaptación del test, 3) administración del test, y 4) interpretación y documentación de las puntuaciones. Hambleton y Patsula (1999) llevan a cabo una sistematización del proceso de adaptación de tests, que se resume (para los no supersticiosos) en trece pasos consecutivos (ver Muñiz y Hambleton, 2000): 1) asegurarse de que existe una equivalencia de los constructos en los idiomas y grupos culturales de interés, 2) decidir si se adapta un test ya existente o se desarrolla uno nuevo, 3) seleccionar traductores profesionales cualificados, 4) combinar diseños de traducción (forward y/o backward), 5) revisar la versión adaptada del test y realizar las correcciones necesarias, 6) llevar a cabo un estudio piloto con el test adaptado, 7) llevar a cabo una prueba empírica rigurosa del test adaptado, 8) elegir un diseño para conectar las puntuaciones de las versiones original y objetivo, 9) si existe interés en hacer comparaciones interculturales hay que asegurarse de la equivalencia de las versiones, 10) llevar a cabo estudios rigurosos de validación, 11) documentar todo el proceso seguido y preparar un manual para los usuarios del test adaptado, 12) entrenar a los usuarios, 13) realizar un seguimiento del test adaptado.

En definitiva, este extenso campo de investigación se centra en estos momentos en tres aspectos clave (Hambleton y Kanjee, 1995): 1) desarrollo de métodos y procedimientos para adaptar tests focalizándose en establecer la equivalencia de puntuaciones, donde juega un papel importante el estudio del DIF y del DTF (Ellis, 1989; Hambleton y Kanjee, 1995; Hidalgo y López, 2000; Muñiz y Hambleton, 2000), 2) diseño de estrategias para interpretar y usar datos transculturales y transnacionales, y 3) desarrollo y uso de directrices o normas para adaptar tests (Hambleton, 1993, 1994a; Hambleton, Merenda y Spielberg, 2001; Muñiz y Hambleton, 1996; Van der Vijver y Hambleton, 1996).

La adaptación/traducción de tests fue uno de los puntos analizados en la reunión de Junio de 1996 por la Comisión Europea de Tests (Muñiz, 1996), que recalcó las implicaciones que tiene en el uso e interpretación de las puntuaciones de los tests; la revista *European Journal of Psychological Assessment* dedica el número 3 del volumen 11 (1995) íntegramente a este

tópico. Además, durante los últimos diez años, el ITC ha estado trabajando en el establecimiento de directrices para el uso (justo y ético) y desarrollo de los tests que incluyen también las pautas recomendadas para una correcta adaptación/traducción de un test (ver Bartram, 2001).

Tests adaptativos informatizados

La elaboración de tests adaptativos informatizados ha sido una de las aplicaciones de la IRT que ha tenido un fuerte impacto en la medida educativa y psicológica, como señala Wainer (2000b), sólo durante el año 1999 se aplicaron a más de un millón de personas. Que el tema es de gran interés se pone de manifiesto por el elevado número de trabajos a este respecto en la última década. Así, en los últimos años han aparecido los libros de Sands, Waters y McBride (1997), Drasgow y Olson (1999) y Wainer (2000a) y el artículo de revisión de Hontangas, Ponsoda, Olea y Abad (2000). El impacto del tema también se deja notar en las revistas psicométricas, con los números monográficos de *Journal of Educational Measurement* (1997, vol. 34, nº 1), *Applied Psychological Measurement* (1999, vol. 23, nº 3), y *Psicológica* (2000, vol. 21, nº 1 y 2).

Un test adaptativo computerizado (Computerized Adaptive Testing, CAT) consiste en un banco de ítems calibrado (almacenado en un ordenador) y un procedimiento de selección de ítems en función de la habilidad de cada sujeto. El uso de tests adaptativos rompe con dos reglas tradicionales de la medida (Embretson y Reise, 2000): 1) para comparar las puntuaciones de los sujetos obtenidas en diferentes formas de un test es necesario trabajar con formas paralelas del test y 2) tests con más ítems son más fiables.

En la TCT se asume que tests más largos presentan una mayor fiabilidad, como queda enunciado en la conocida fórmula de la profecía de Spearman-Brown. Lo anterior es cierto siempre que se trate de tests de un tamaño fijo, y se cumple tanto para la TCT como para la IRT que el error de medida es mayor para un test con menor número de ítems que para otro con mayor número. Sin embargo, un test adaptativo, en el que los ítems son seleccionados para cada sujeto en función de su nivel en el rasgo, conlleva un mayor nivel de precisión que el correspondiente test fijo de mayor tamaño. El uso de test adaptativo presupone que sujetos con diferentes niveles del rasgo se les administra diferentes ítems. Si el banco de ítems es lo suficientemente amplio, es posible obtener el mismo error típico independientemente del nivel de habilidad de los sujetos.

Conclusiones

Se han presentado las teorías de medida, IRT y GT, más actuales en Psicometría y se han comentado sus aportaciones más ventajosas con respecto a la TCT. En todo este desarrollo se ha prestado más atención a la IRT por ser uno de los enfoques no sólo de mayor crecimiento e

impacto científico, sino por las atractivas y poderosas consecuencias de su uso desde el punto de vista de la construcción de tests y la interpretación de las puntuaciones que se obtienen. Sin embargo, en la presentación de la IRT se ha evitado entrar en algunas de sus polémicas y por supuesto de sus limitaciones, que ahora se tratarán de abordar.

En primer lugar, una de las limitaciones importantes de la IRT es que los supuestos en los que se basa son muy restrictivos, partiendo de supuestos acerca de los datos que en situaciones aplicadas no siempre se cumplen. Por ejemplo, el supuesto de independencia local de ítems en algunos datos es de difícil cumplimiento, lo que ha llevado al desarrollo de modelos de IRT menos restrictivos donde se asume dependencia local de los ítems (ver Mellenbergh, 1994 y Van der Linden y Hambleton, 1997 para una revisión de éstos y otros modelos de IRT). Por otro lado, el supuesto de unidimensionalidad exigible para ajustar algunos de los modelos de IRT, tampoco se cumple en la práctica; aunque las soluciones pasan por ajustar modelos multidimensionales, éstos se encuentran en sus primeros desarrollos, y tanto la estimación como la interpretación de los parámetros que se obtienen de su uso presentan dificultades.

En segundo lugar, la aplicación de un modelo de respuesta al ítem requiere de un elevado tamaño muestral para asegurar que las estimaciones de los parámetros son estables. Este problema se incrementa cuando se trabaja con ítems de respuesta politómica, donde al menos se requiere un tamaño muestral de cinco veces más que el número de ítems por el de categorías del ítem.

En tercer lugar, los modelos de respuesta al ítem multiparamétricos presentan problemas para establecer una medida objetiva (Fisher, 1994; Wright, 1977). Existe un fuerte debate entre los defensores del modelo de Rasch y los partidarios de los modelos con múltiples parámetros. Para los primeros, el modelo uniparamétrico es el único que proporciona una medida fundamental y cumple la propiedad de especificidad objetiva, proporcionando estimaciones suficientes de los parámetros. Con respecto a los segundos, Hambleton (1994b) hace una exposición detallada de las evidencias prácticas y teóricas que avalan el uso de modelos multiparamétricos sobre la base de su significado, bondad de ajuste y utilidad. Lord (1980) señala a este respecto que no sólo es importante el concepto de suficiencia en las estimaciones de los parámetros sino también el de información.

En resumen, aunque los supuestos en los que se basa la IRT son supuestos “fuertes”, frente a los “débiles” en los que descansa la TCT, el lector no debe confundirse y pensar que la aplicación de estos modelos a los datos de un test va a proporcionar resultados “fuertes”, esto último se obtendrá si los datos con los que se trabaja cumplen los supuestos del modelo que se trata de ajustar. Es más, ningún modelo de medida (más o menos potente en sus consecuencias) puede compensar un inadecuado diseño de recogida de los datos ni la falta de datos.

En la última parte de este trabajo se han abordado temas relacionados con el uso y aplicación de los tests. Así, se comentan algunas de las posturas más recientes sobre la problemática del funcionamiento diferencial de los tests y de los ítems, la importancia de utilizar una metodología adecuada cuando se adaptan o traducen tests de una cultura a otra y el creciente impacto de los tests adaptativos computerizados. Estos temas son de actualidad en revistas especializadas en psicometría, pero también lo son en revistas aplicadas de psicología, educación y medicina, principalmente desde su perspectiva más práctica.

Es más, tanto los teóricos como los prácticos andan en estos momentos muy preocupados por el buen uso de los tests y por la calidad de los mismos, destacando un creciente interés en establecer normas y directrices que regulen el uso de los tests. En este sentido las asociaciones más fuertes del ámbito psicológico, educativo y de la medida (APA, AERA, NCME, ITC) se encuentran ocupadas, entre otros menesteres, en definir reglas éticas para el uso de los tests y para la construcción de los mismos con la finalidad de evitar la presencia de sesgos y factores culturales, la injusticia en la obtención de puntuaciones y, por supuesto, en la toma de decisiones. Todas estas asociaciones destacan la importancia de conocer a fondo y estar suficientemente formado en los conceptos claves de la medida mediante tests y en los nuevos modelos de medida. En definitiva, la formación adecuada en medida es el eje esencial para que los prácticos que usan tests, los utilicen de forma pertinente y ética.

Referencias

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Anguera, M.T. (1990). Metodología observacional. En J. Arnau, M.T. Anguera y J. Gómez, *Metodología de la investigación en ciencias del comportamiento*. Murcia: Secretariado de Publicaciones Universidad de Murcia.

Baker, F.B. (1985). *The basis of ítem response theory*. Portsmouth, NH: Heinemann.

Baker, F.B. (1987). Item parameter estimation under the one, two and three parameter logistic model. *Applied Psychological Measurement, 11,* 111-141.

Baker, F.B. (1992). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker.

Barbero, I. (1999). Desarrollos recientes de los modelos psicométricos de la Teoría de Respuesta a los ítems. *Psicothema, 11,* 195-210.

Bartram, D. (2001). The Development of International Guidelines on Test Use: The International Test Commission Project. *International Journal of Testing,* 1, 33-53.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En F.M. Lord y M. Novick, *Statistical theories of mental scores*. Reading, MA: Addison-Wesley.

Blanco, A. (1989). Fiabilidad y generalización de la observación conductual. *Anuario de Psicología, 43*, 5-32.

Blanco, A. (2001). Generalizabilidad de observaciones uni y multifaceta: estimadores LS y ML. *Metodología de las Ciencias del Comportamiento*, 3 (2), 161-198.

Bock, R.D. (1997). A brief history of Item Response Theory. *Educational Measurement: Issues and Practice, 16,* 21-33.

Bock, R.D. y Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*, 179-197.

Brennan, R.L. (1983). *Elements of generalizability theory*. Iowa City: ACT Publishers.

Brennan, R.L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*, 27-34.

Brennan, R.L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice, 16*, 14-20.

Brennan, R.L. (2000). Performance assessment from the perspective of generalizability theory. *Applied Psychological Measurement, 24*, 339-353.

Brennan, R.L., Gao, X. y Colton, D.A. (1995). Generalizability analyses of word keys listening and writing tests. *Educational and Psychological Measurement, 56*, 157-176.

Cella, D. y Chang, C.H. (2000). A discussion of ítem response theory and its applications in health status assessment. *Medical Care, 38 (Suppl. 9)*, II66-II72.

Cronbach, L.J., Gleser, G.C., Nanda, H. y Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, L.J., Linn, R.L., Brennan, R.L. y Haertel, E. (1997). Generalizability analysis for performance assessments of students achievement or school efectiveness. *Educational and Psychological Measurement, 57*, 373-399.

Drasgow, F. y Olson, J.B. (Eds.) (1999). *Innovations in computerized assessment*. New Jersey: LEA.

Ellis, B.B. (1989). Differential ítem functioning: Implications for test translation. *Journal of Applied Psychology, 74*, 912-921.

Embretson, S.E. (1996). The new rules of measurement. *Psychological Assessment, 8*, 341-349.

Embretson, S.E. y Prenovost, L.K. (1999). Item response theory in assessment research. En P.C. Kendall, J.N. Butcher, y G.N. Holmbeck (Eds.), *Handbook of research methods in clinical psychology* (2nd ed.). New York, NY: John Wiley & Sons.

Embretson, S.E. y Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ, US: Lawrence Erlbaum Associates.

Engelhard, G. (1992). Historical views of the concept of invariance in measurement theory. En M. Wilson (Ed.), *Objective measurement: Theory into practice (Vol. 1)*. Norwood, NJ: Ablex Publishing Corporation.

Engelhard, G. y Wilson, M. (Eds.) (1996). *Objective measurement: Theory into practice (Vol. 3)*. Norwood, NJ: Ablex Publishing Corporation.

Feldt, L.S. y Brennan, R.L. (1989). Reliability. En R.L. Linn (Ed.), *Educational Measurement* (3th Edition) (pp. 105-146). New York: American Council on Education and Macmillan publishing company.

Feldt, L.S., Steffen, M. y Gupta, N.C. (1985). A comparison of five methods for estimating standard error of measurement at specifid score levels. *Applied Psychological Measurement*, 9, 351-361.

Fischer, G.H. y Molenaar, I.W. (1995). *Rasch models. Foundations, recent developments and applications*. New York: Springer Verlarg.

Fisher, W.P. (1994). The Rasch debate: Validity and revolution in educational measurement. En M. Wilson (Ed.), *Objective measurement: Theory into practice. (Vol. 2)*. Norwood, NJ: Ablex Publishing Corporation.

Fitzpatrick, A.R., Guemin, L. y Furong, G. (2001). Assessing the comparability of school scores across test forms that are not parallel. *Applied Measurement in Education, 14*, 285-306.

Fidalgo, A.M. (1996). Funcionamiento diferencial de los items. En J. Muñiz (Coord.), *Psicometría*. Madrid: Universitas.

Fleer, P.F., Raju, N.S. y van der Linden, W.J. (1995). *A Monte Carlo assessment of DFIT with dichotomously scored unidimensional tests*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Flowers, C.P., Oshima, C. y Raju, N.S. (1996). *A description and demostration of the polytomously-DFIT framework*. Unpublished paper.

Flowers, C.P., Oshima, C. y Raju, N.S. (1997). *The relationship between polytomously DFIT and other polytomous DIF procedures*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Gillmore, G.M. (1983). Generalizability theory: applications to program evaluation. En L.J. Fyans (Ed.), *New directions for testing and measurement: Generalizability theory: Inferences and practical application. (Nº 18)*. San Francisco: Jossey-Bass.

Goldstein, H. y Wood, R. (1989). Five decades of ítem response modelling. *British Journal of Mathematical and Statistical Psychology, 42*, 139-167.

Gómez, J. e Hidalgo, M.D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: Una revisión metodológica. *Anuario de Psicología, 74*, 3-32.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139-150.

Hambleton, R.K. (1990). Item response theory: Introduction and bibliography. *Psicothema, 2*, 97-107.

Hambleton, R.K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment*, 9, 54-65.

Hambleton, R.K. (1994a). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*, 229-244.

Hambleton, R.K. (1994b). Item response theory: A broad psychometric framework for measurement advances. *Psicothema, 6*, 535-556.

Hambleton, R.K. y Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment*, 11, 147-157.

Hambleton, R.K., Merenda, P., y Spielberger, C. (Eds.). (2001). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: LEA.

Hambleton, R.K. y Patsula, L. (1999). Increasing the validity of Adapted Tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, vol. 1*, http://www. Testpublishers.org/journal1.htm

Hambleton, R.K. y Swaminathan, J. (1985). *Item response theory: principles and applications*. Boston, MA: Kluwer-Nijhoff Publishing.

Harvey, R.J. y Hammer, A.L. (1999). Item response theory. *Counseling Psychologist, 27*, 353-383.

Hays, R.D., Morales, L.S. y Reise, S.P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38* (suppl.9), II28-II42.

Herrera, A. N., Sánchez, R. y Gómez, J. Funcionamiento diferencial de los ítems: Una revisión conceptual y metodológica. *Acta Colombiana de Psicología*, 5, 41-61.

Hidalgo, M.D. y Gómez, J. (1999). Técnicas de detección del funcionamiento diferencial en ítems politómicos. *Metodología de las Ciencias del Comportamiento, 1*, 39-60.

Hidalgo, M.D. y López, J.A. (2000). Funcionamiento diferencial de los ítems: Presente y perspectivas de futuro. *Metodología de las Ciencias del Comportamiento, 2*, 167-182.

Holland, P.W. y Rubin, D.B. (Eds.) (1982). *Test equating*. New York: Academic Press.

Holland, P.W. y Wainer, H. (Eds.) (1993). *Differential Item Functioning*. Hillsdale, NJ: LEA.

Hontangas, P., Ponsoda, V., Olea, J. y Abad, F. (2000). Los tests adaptativos informatizados en la frontera del siglo XXI: una revisión. *Metodología de las Ciencias del Comportamiento*, 2, 183-216.

Hulin, C.L., Drasgow, F. y Parsons, C.K. (1983). *Item response theory: Applications to psychological measurement*. Homewood, IL: Dow-Jones-Irwin.

Jones, L.V. y Appelbaum, M.I. (1989). Psychometrics Methods. *Annual Review of Psychology, 40*, 23-43.

Kolen, M.J. y Brennan, R.L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlarg.

Lawley, D.N. (1943). On problems connected with ítem selection and test construction. *Proceedings of the Royal Society of Edimburg, 61*, 273-287.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. En S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star y J.A. Clausen (Eds.), *Measurement and prediction*. Princeton, NJ: Princeton University.

Levine, M.V. y Rubin, D.B. (1979). Measuring appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269-290.

López Pina, J.A. (1995). *Teoría de Respuesta al ítem: Fundamentos.* Barcelona: PPU.

López Pina, J.A. e Hidalgo, M.D. (1996). Bondad de ajuste y teoría de la respuesta a los ítems. En J. Muñiz (Coord.), *Psicometría*. Madrid: Universitas.

Lord, F.M. (1952). A theory of test scores. *Psychometric Monographs, 7*.

Lord, F.M. (1980). *Applications of ítem response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F.M. y Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Marcoulides, G.A. (1989). The application of generalizability analysis to observational studies. *Quality & Quantity, 23*, 115-127.

Meijer, R.R. y Nering, M.L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement, 21*, 321-326.

Meijer, R.R. y Sijtsma, K. (1995). Detection of aberrant ítem score patterns: A review of recent development. *Applied Measurement in Education, 8*, 261-272.

Mellenbergh, G.J. (1994). Generalized linear ítem response theory. *Psychological Bulletin, 115*, 300-307.

Millsap, R.E. y Everson, H.T. (1993). Methodology Review: Statistical approaches for assesing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

Muñiz, J. (1990). *Teoría de Respuesta a los items*. Madrid: Pirámide.

Muñiz, J. (1996). Reunión en Madrid de la Comisión Europea de Tests. *Papeles del Psicólogo, 65*, 89-91.

Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Items*. Madrid: Pirámide.

Muñiz, J. y Hambleton, R.K. (1992). Medio siglo de teoría de respuesta a los ítems. *Anuario de Psicología*, 52, 41-66.

Muñiz, J. y Hambleton, R.K. (1996). Directrices para la traducción y adaptación de los tests. *Papeles del Psicólogo, 66*, 63-70.

Muñiz, J. y Hambleton, R.K. (2000). Adaptación de los tests de unas culturas a otras. *Metodología de las Ciencias del Comportamiento, 2*, 129-149.

Navas, M.J. (1996). Equiparación. En J. Muñiz (Coord.) *Psicometría*. Madrid: Universitas.

Navas, M.J. (2000). Equiparación de puntuaciones: exigencias actuales y retos de cara al futuro. *Metodología de las Ciencias del Comportamiento*, 2, 151-165.

Oakland, T. y Hu, S. (1992). The top ten tests used with children and youth worldwide. *Bulletin of the International Test Commission, 19*, 99-120.

Oakland, T., Poortinga, Y.H., Schlegel, J. y Hambleton, R.K. (2001). International Test Commission: Its History, Current Status, and Future Directions. *International Journal of Testing, 1*, 3-32.

Oshima, T.C., Raju, N.S., Flowers, C. y Monaco, M. (1995). *A Montecarlo assessment of DFIT with dichotomously scored multidimensional tests*. Paper presented at the Annual Meeting of American Educational Research Association, San Francisco.

Potenza, M.T. y Dorans, N.J. (1995). DIF Assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, 19, 23-37.

Peterson, N., Marco, G. y Steward, E. (1982). A test of the adequacy of linear score equating models. En P. Holland y D. Rubin (Eds.), *Test equating*. New York: Academic Press.

Raju, N.S., van der Linden, W.J. y Fleer, P.F. (1995). IRT-based internal measures of differential ítem functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368.

Rasch, G. (1960). *Probabilistic Models for some Intellingence and Attainment Tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut, (Chicago, IL: University of Chicago Press, 1980).

Reise, S.P. y Waller, N.G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology, 65*, 143-151.

Richardson, M.W. (1936). The relationship between difficulty and the differential validity of a test. *Psychometrika, 1*, 33-49.

Rozeboom, W.W. (1978). Domain validity why care?. *Educational and Psychological Measurement, 38*, 81-88.

Sands, W.A.; Waters, B.K. y McBride, J.R. (Eds.) (1997). *Computer adaptive testing: Form Inquiry to operation*. Washington, DC: American Psychological Association.

Shavelson, R.J. y Webb, N.M. (1991). *Generalizability Theory. A primer*. London: Sage.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

Traub, R.E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16 (4)*, 8-14.

Traub, R.E. y Lam, Y.R. (1985). Latent structure and ítem sampling models for testing. *Annual Review of Psychology, 36*, 19-48,

Tucker, L.R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1-13.

Van der Linden, W.J. y Hambleton, R.K. (1997). *Handbook of modern ítem response theory*. New York: Springer Verlag.

Van der Linden, W.J. y Zwart, M.A. (1989). Some procedures for computerized ability testing. *International Journal of Educational Research, 13*, 175-188.

Van de Vijver, F.J.R. y Hambleton, R.K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89-99.

Wainer, H. (2000a). CATs: Whiter and whence. *Psicológica, 21*, 121-133.

Wainer, H. (Ed.) (2000b). *Computerized adaptive testing: A primer* (2nd. Ed.). Princeton, NJ: Educational Testing Service.

Weiss, D.J. (Ed.) (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.

Wilson, M. (Ed.) (1992), *Objective measurement: Theory into practice (Vol. 1)*. Norwood, NJ: Ablex Publishing Corporation.

Wilson, M. (Ed.) (1994), *Objective measurement: Theory into practice (Vol. 2)*. Norwood, NJ: Ablex Publishing Corporation.

Wilson, M. y Engelhard, G. (Eds.) (En prensa), *Objective measurement: Theory into practice (Vol. 5)*. Norwood, NJ: Ablex Publishing Corporation.

Wilson, M., Engelhard, G. y Draney, K. (Eds.) (1997), *Objective measurement: Theory into practice (Vol. 4)*. Norwood, NJ: Ablex Publishing Corporation.

Wood, R.L., Wingersky, M.S. y Lord, F.M. (1976). *LOGIST: A computer program for estimating examinee ability and ítem characteristic curve parameters*. (Research Memorandum, 76-6). Princeton, NJ: Educational Testing Service.

Woodruff, D. (1990). Conditional standard error of measurement in prediction. *Journal of Educational Measurement, 14*, 97-116.

Wright, B.D. (1977). Solving measurements problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.

Wright, B.D. y Mead, R.J. (1976). *BICAL: Calibrating items with the Rasch model*. (Research Memorandum, 23). Statistical Laboratory, Department of Education, University of Chicago.

Wright, B.D. y Stone, M.H. (1979). *Best test design*. Chicago IL: Mesa Press.

Yen, M. y Edwardson, S.R. (1999). Item response theory approach in scale development. *Nursing Research, 48*, 234-238.

Zeng, L. (1993). A numerical approach for computing standard errors of linear equating. *Applied Psychological Measurement, 18*, 369-378.

Zeng, L. y Cope, R.T. (1995). Standard error of linear equating for the conterbalanced design. *Journal of Educational and Behavioral Statistics, 20*, 447-478.

Autores

Juana Gómez Benito; Departamento de Metodología de las Ciencias del Comportamiento; Facultad de Psicología; Universidad de Barcelona. Passeig Vall d'Hebrón, 171; 08035 Barcelona, España. E-mail: jgomez@psi.ub.es

María Dolores Hidalgo Montesinos. Facultad de Psicología Universidad de Murcia 30080 Murcia, España. E-mail: mdhidalg@um.es