Psicothema, 1999. Vol. 11, nº 1, pp. 195-210
ISSN 0214 - 9915 CODEN PSOTEG

DESARROLLOS RECIENTES DE LOS MODELOS PSICOMÉTRICOS DE LA TEORÍA DE RESPUESTA A LOS ÍTEMS

Mª Isabel Barbero García
UNED

Desde que Lord (1952) publicara su trabajo titulado *A theory of test scores* hasta nuestros días, los modelos de la TRI han ido introduciendose con fuerza en el campo de la evaluación psicológica y educativa, como lo demuestran las investigaciones recogidas en distintas revistas especializadas y presentadas en reuniones científicas. Se han hecho también excelentes trabajos de recopilación que ponen de manifiesto la evolución de estos modelos y los problemas prácticos que han ido solucionando, uno de ellos es el llevado a cabo por van der Linden y Hambleton (1997) que incluye importantes trabajos dentro del campo de la psicometría. El objetivo de nuestro trabajo ha sido llevar a cabo una síntesis de los modelos allí presentados pero de una manera sencilla, sin el aparato formal, de manera que todas aquellas personas interesadas en el tema puedan tener una visión general del mismo y, más tarde, puedan abordar el estudio de estos modelos con mayor profundidad.

Recent developments of irt psychometric models. Since the publication of Lord's (1952) *A ttheory of test scores* until now, large numbers of item response models have been introduced in the psychological and educational field. This fact has implied a great number of researches and scientific papers presented at international meetings or published in specialized journals. There have also been some excellent compilations about the evolution of these models as well as the problems that these models have solved. One of these compilations is the one carried out by van der Linden and Hambleton (1997), who have collected relevant works in the psychometric field. The aim of our work is to carry out a general review of these models presented by van der Linden and hambleton (1997), without their formal framework and in a simple way, so that people interested in the topic can have a general point of view of the actual state of the art.

Durante la última década han proliferado considerablemente el número de trabajos realizados en España dentro del marco de la Teoría de Respuesta al Ítem (TRI), trabajos que abarcan tanto su vertiente teórica como aplicada. Próximamente aparecerá un trabajo en el que se recogen, agrupadas por categorías temáticas, las principales referencias de los artículos publicados y de los trabajos presentados a los distintos congresos y reuniones científicas, de forma similar al trabajo realizado por Hambleton (1990) con los publicados en Estados Unidos; de esta forma se pondrá de manifiesto la importancia

Correspondencia: Mª Isabel Barbero García
Facultad de Psicología. UNED
Ciudad Universitaria, s/n.
28040 Madrid (Spain)
E mail: ibarbero@cu.uned.es

de las investigaciones llevadas a cabo en nuestro país dentro del campo de la TRI. Los trabajos citados a continuación pueden servir de ejemplo, en ellos puede apreciarse la consolidación de algunos equipos de investigación: Barbero y Prieto (1995, 1997), Barbero y Navas (1995), Barbero (1996, 1997), Cuesta y Muñiz (1994, 1995), Cuesta (1996), Ferrando (1996), Fidalgo (1996), Gómez e Hidalgo (1997), Gómez y Navas (1996), Hidalgo y López (1997), López- Pina (1995), López-Pina e Hidalgo (1996), Martinez-Cardeñoso, Cuesta y Muñiz (1996), Maydeu (1996), Molina (1997), Muñiz, Rogers y Swaminathan (1989), Muñiz (1990, 1996), Muñiz y Hambleton (1992), Navas (1993, 1994, 1996), Olea y Ponsoda (1996), Ponsoda, Olea y Revuelta (1994), Prieto y Barbero (1996, 1997), Prieto y Delgado (1996), Renom (1993, 1998), San Luis, C., Prieto, P., Sánchez-Bruno, A. y Barbero, I. (1995).

Se han hecho muy buenas introducciones a la TRI (Muñiz, 1990, 1997; López Pina, 1995 entre otros). Una revisión completa de las ventajas e inconvenientes de la Teoría de Respuesta al Ítem respecto a la Teoría Clásica de los Tests puede encontrarse en Navas (1994) y una revisión de las principales aportaciones de la Teoría de Respuesta al Ítem desde sus orígenes en Muñiz y Hambleton (1992).

Teniendo en cuenta que en los últimos años se ha asistido al desarrollo de nuevos modelos psicométricos de TRI, a lo largo de estas páginas se intentará dar una visión amplia de los mismos y de los problemas prácticos que han venido a solucionar, teniendo en cuenta que ocupan, hoy día, un lugar principal en la literatura psicométrica, y que su campo de aplicación se ha extendido a muchas áreas de la psicología.

Todo lo dicho en ellas ya se había escrito con anterioridad, van der Linden y Hambleton (1997) hacen una recopilación completa de lo que aquí se expone, el objetivo que ha guiado el trabajo ha sido hacer una síntesis que sirva de introducción a todas aquellas personas interesadas en el tema y descubrirles las posibilidades de investigar en los nuevos modelos y sus aplicaciones.

La variedad de modelos que ha ido desarrollándose a lo largo del tiempo dificulta la organización de los mismos para su revisión. Han sido muchos los intentos realizados al respecto (Masters y Wright, 1984 y Thissen y Steinberg, 1984 entre otros); sin embargo, a la hora de realizar este trabajo se ha seguido la clasificación de van der Linden y Hambleton (1997) por considerar que era la que más se ajustaba a nuestro propósito.

Asumiendo que el objetivo central de la TRI es la especificación de una función matemática que relacione la probabilidad de que un sujeto responda correctamente a un ítem cualquiera con una característica o rasgo subyacente denominado θ, se observa un cambio sustancial respecto a la Teoría Clásica de los Tests (TCT) en la medida en que el interés de la TRI está centrado en la actuación de los sujetos en cada uno de los ítems y no en el test total. Esta función ha recibido varios nombres aunque, hoy día, se la conoce con el nombre de *Función Característica del Ítem (FCI), o Curva Característica del Ítem (CCI)* y las distintas formas que pueda tomar dará lugar a distintos modelos.

Modelos unidimensionales para ítems dicotómicos: Ojiva normal y Logísticos

Haciendo un poco historia, se puede decir que el primer modelo que se desarrolló fue el modelo de *Ojiva normal* que, aunque se atribuye a Lawley (1943), fue estudiado con anterioridad por Ferguson (1942), Mosier (1940, 1941) y Richardson (1936) entre otros. La idea original fue utilizada por Thurstone en 1927 al asumir entre los supuestos de su modelo que la distribución de los valores asignados por los sujetos a los estímulos, a través de los distintos procesos

discriminativos suscitados ante la presentación de los mismos, era una distribución normal.

La terminología utilizada en los trabajos de estos autores refleja la influencia de los estudios de Psicofísica, cuyo objetivo era estudiar las relaciones entre las propiedades físicas de los estímulos y las sensaciones que suscitaban en los sujetos. El método más utilizado consistía en presentar un estímulo que variaba en intensidad y el sujeto debía responder si detectaba o no el estímulo. Teniendo en cuenta que la probabilidad de detección es una función creciente de la intensidad física del estímulo, se utilizó la función normal como función de respuesta. En esta función, el parámetro θ era conocido y correspondía a la intensidad del estímulo, por lo tanto el interés estaba centrado fundamentalmente en los parámetros a_i y b_i,este último conocido como *umbral* de los estímulos (van der Linden y Hambleton, 1997).

El uso de la distribución normal estaba tan arraigado que incluso si una variable no se distribuía conforme a dicha distribución se normalizaba para poder asumir el modelo.

Aunque Lord (1952) en su libro *A Theory of Test Scores* presenta el primer tratamiento coherente del modelo de ojiva normal, es en la década de los 60 cuando la TRI comienza su gran desarrollo con la publicación del trabajo de Rasch (1960) y la aparición del libro de Lord y Novick (1968) *Statistical theories of mental test scores* en el que se incluye un trabajo de Birnbaum. Estos dos trabajos marcarán la aparición de otros modelos: *los Modelos logísticos para datos dicotómicos* que son los que han tenido y siguen teniendo una mayor influencia en la TRI tanto a nivel práctico como desde el punto de vista histórico. En estos modelos se asume que la relación entre la respuesta de un sujeto a un ítem determinado y la característica o rasgo latente que mide, puede ser descrita por la función de distribución logística.

Rasch comenzó sus trabajos en el campo de la medición educativa y psicológica a finales de 1940, y a comienzos de la década de los 50 desarrolló dos modelos de Poisson para tests de lectura y un modelo para tests de rendimiento e inteligencia al que se conoce como *modelo de Rasch* y es un modelo logístico de un parámetro. Formalmente, el modelo es un caso especial del modelo de Birnbaum, aunque suelen estudiarse de forma separada por las características propias que tiene.

El objetivo de Rasch al desarrollar su modelo era tratar de eliminar la dependencia que existía entre los parámetros de los ítems y las características de la muestra de sujetos utilizada; para Rasch, sólo merecería la pena hacer los análisis de los tests si estuvieran centrados en los sujetos y tuvieran parámetros independientes para los ítems y para los sujetos.

Este punto de vista marcó la transición de la TCT, centrada en la aleatorización y estandarización y en la que los parámetros de los ítems dependían de la muestra utilizada, a la TRI con sus modelos probabilísticos que permiten establecer una relación funcional entre cada ítem y cada sujeto. La contribución de Rasch al problema de la medición fue decisiva al darse cuenta de las posibilidades que ofrecía el modelo logístico de un parámetro para poder hacer estimaciones independientes para los parámetros de los ítems y para el nivel de habilidad de los sujetos.

A diferencia de Rasch, los trabajos de Birnbaum no tenían como objetivo la elaboración de una nueva teoría de tests mediante el desarrollo de un modelo sino hacer posibles, a nivel estadístico, los trabajos comenzados por Lord (1952) y encontrar el modelo o modelos que mejor se ajustaran a los datos obtenidos con la aplicación de los tests. En particular, proporcionó la teoría estadística para la estimación de los parámetros de los ítems y el de habilidad de los sujetos y,

entre otras cosas, propuso una aproximación racional a la construcción de tests; sin embargo, la contribución más importante fue la sugerencia de sustituir el modelo de ojiva normal por el modelo logístico, mucho más fácil de tratar. Otra de sus contribuciones fue la propuesta de incluir en el modelo logístico un tercer parámetro que justificara la obtención de puntuaciones distintas de cero en los sujetos del nivel más bajo en habilidad cuando los ítems son dicotómicos o de elección múltiple, debido a la posibilidad de acierto por azar. Este modelo es conocido como el *modelo logístico de tres parámetros* a diferencia del anterior que se conocía como *modelo logístico de dos parámetros*.

Ni Lord ni Birnbaum mostraron ningún interés por los modelos de un parámetro, ni el de ojiva normal ni el logístico, ya que creían que los modelos necesitaban al menos dos parámetros para representar a los ítems, uno para la dificultad del ítem y otro para su poder discriminativo; sin embargo, el modelo logístico de un parámetro, formalmente, es un caso especial de los modelos que ellos mismos desarrollaron.

A partir de la publicación de los trabajos de Rasch (1960) y Birnbaum (1968), comienzan a realizarse numerosas aplicaciones prácticas de los modelos y se va extendiendo su campo de aplicación, pero hasta los años 80 la mayoría de las investigaciones estuvieron centradas en problemas de estimación de parámetros, bondad de ajuste, robustez de los modelos y aplicaciones de estos modelos; pocas investigaciones estaban encaminadas al desarrollo de nuevos modelos.

Es a partir de los años 80 cuando la TRI cobra su verdadero sentido, al poder disponer del soporte matemático, informático y tecnológico para su aplicación y desde ese momento el uso de los modelos se empieza a generalizar tanto en Empresas, como en Departamentos de Educación, en las Fuerzas Armadas, etc. y se llevan a cabo aplicaciones a gran escala en las que los datos se analizan mediante estos modelos.

Tanto los modelos de ojiva normal como los modelos logísticos que hemos expuesto se desarrollaron para el análisis de ítems con formato de respuesta dicotómico y asumen dos principios o supuestos: El primero es el supuesto de *unidimensionalidad* que implica que la respuesta de los sujetos a cada uno de los ítems, y por lo tanto su actuación en el test, depende de su nivel en una única característica, rasgo o dimensión, designado por θ. El segundo, es el supuesto conocido como de *independencia local,* que postula la independencia estadística entre las respuestas dadas por los sujetos a cada uno de los ítems, de manera que la respuesta dada a uno de ellos no influye en las respuestas que den a cualquiera de los otros.

Hoy día existen numerosos programas de ordenador que facilitan la utilización práctica de estos modelos: LOGIST, BILOG, BICAL, MICROCAT, METRIX, ANATRI, etc.

En la práctica, sin embargo, se ha puesto de manifiesto la necesidad de utilizar ítems con un formato de respuesta menos restrictivo: nominal, de crédito parcial, de respuesta graduada, politómica, de respuesta abierta, de respuesta construida. También puede ser necesario construir pruebas que, además de las habilidades básicas, midan habilidades o características multidimensionales, actitudes en lugar de habilidades, rasgos de personalidad, o ser diseñados para diagnosticar procesos cognitivos subyacentes. Estos problemas no podían solucionarlos los modelos de ojiva normal y logística conocidos hasta el momento y, por eso, se empezaron a desarrollar otros que cubrieran esas necesidades. Aunque algunos de ellos se publicaron a finales de los 60 y principios de los 70 (Bock, 1972; Rasch, 1961; Samejima, 1969, 1972), la mayoría de ellos se desarrollaron a partir de los años 80 que, como hemos comentado, es la fecha que marca el gran desarrollo de la TRI.

Mellenbergh (1994) mostró que la mayoría de los modelos pueden ser obtenidos a partir del modelo lineal generalizado y propuso un marco de referencia al que denominó GLIRT (*Generalized Linear Item Response Theory).* Esta teoría es válida para la mayoría de los modelos que se conocen hoy día.

Modelos unidimensionales para ítems politómicos

A finales de los años 60 y comienzos de los 70 comienzan una serie de investigaciones basadas en la idea de que si las distintas alternativas que hay en un ítem (distractores) juegan un papel importante en el proceso de respuesta y permiten obtener un mayor grado de información acerca del nivel de habilidad de los sujetos, será más útil que en lugar de puntuar las respuestas a los ítems como correctas o incorrectas se tengan en cuenta y se puntúen todas y cada una de las respuestas dadas. Si esto es así, las propiedades de estas alternativas de respuesta deben ser parametrizadas en el modelo. El pionero en este área fue Fumiko Samejima quien hacia la mitad de los 60 desarrolló una serie de *Modelos para Respuestas Graduadas* que utilizaban la ojiva normal y la logística, y podían ser aplicados a datos de categorías ordenadas, como los procedentes de escalas tipo Likert. Estos modelos incluyen funciones de respuesta para cada una de las alternativas. Su trabajo en este área estuvo motivado, tanto por el deseo de generar e investigar nuevos modelos, como por la utilización cada vez mayor de pruebas con ítems politómicos en el campo de la evaluación educativa. A partir de estos modelos se desarrollaron numerosas variaciones por ejemplo el *Steps Model* (Verhelst, Glas y de Vries) para analizar el crédito parcial; el *Modelo Secuencial para Respuestas Ordenadas* (Tutz) desarrollado especialmente para aquellos ítems que han de solucionarse paso a paso de forma secuencial ; el *Modelo de Crédito Parcial Generalizado* (Muraki) y el *Modelo de Crédito Parcial* (Masters y Wright) que se diferencia del de Samejima en que pertenece a la familia de modelos de Rasch y es el más sencillo de todos los modelos para categorías ordenadas. Todos ellos intentan modelar las respuestas de los sujetos a ítems con categorías de respuesta politómica ordenadas. Las principales diferencias entre ellos están en la forma de modelización y en el número de parámetros. Existen hoy día algunos programas de ordenador que permiten el análisis del crédito parcial. Los primeros programas para implementar los modelos de crédito parcial fueron el CREDIT (de la Universidad de Chicago), PC-CREDIT (Universidad de Melbourne), y PARCIAL (De la Universidad de Texas en Austin). Hoy día hay un gran número de programas que permiten la implementación de análisis de crédito parcial: QUEST (Adams y Khoo, 1992) y BIGSTEPS (Wright y Linacre, 1992).

Otra variación del modelo de respuesta graduada de Samejima (1968) es el *Modelo Rating Scale* (Andersen) desarrollado como una extensión del modelo de Rasch para datos politómicos, pero asume que las categorías de respuesta son equidistantes de ahí que las puntuaciones asignadas a las categorías deban estar igualmente espaciadas.

Para datos categóricos no ordenados Bock (1972) desarrolló su *Modelo de Respuesta Nominal.* Una extensión del modelo, en la que se trata del problema de las respuestas por azar en los ítems de elección múltiple, es la desarrollada por Thissen y Steinberg con su *Modelo de Respuesta para ítems de elección múltiple.*

El programa MULTILOG de Thissen (1991) y el PARSCALE de Muraki y Bock (1991) permiten la estimación de los parámetros de la mayoría de estos modelos. El programa TESTGRAF (Ramsay, 1992) permite el análisis gráfico de los datos en tests y cuestionarios de elección múltiple.

Modelos para tests con tiempo límite o ítems de ensayo múltiple

Dentro de la teoría de los tests es necesario diferenciar los tests de velocidad de los de potencia, en los primeros la dificultad de los ítems es prácticamente nula y todos podrían ser resueltos correctamente si no se impusiera la restricción de la limitación del tiempo. En los tests de potencia, por el contrario, los ítems tienen dificultad creciente, no hay tiempo límite de aplicación y se asume que el sujeto responde correctamente a aquellas preguntas que conoce. La mayoría de los tests, en la práctica son una mezcla de tests de velocidad y de potencia pues se trata de tests cuyos ítems varían en dificultad y se administran con tiempo límite.

La TRI esencialmente asume que el test es de potencia puro, los sujetos intentan responder a todos los ítems y en el modelo sólo se tienen en cuenta las respuesta correctas ignorándose el tiempo que el sujeto ha tardado en responder.

Los principios de la TRI pueden ser adaptados a los tiempos de respuesta en los tests de velocidad puros y así, en lugar de asumir que la probabilidad de responder correctamente a un ítem es función de los parámetros del ítem y de la habilidad de los sujetos, se asumiría que la probabilidad de que un sujeto tarde un determinado tiempo de responder a un ítem es función de la dificultad del ítem y de la velocidad (mental) del sujeto.

El problema radica, como señalan van der Linden y Hambleton (1997), en que en la práctica, los tests no son de velocidad puros ni de potencia puros, lo que trae algunas consecuencias:

– Si los tests no son de velocidad puros el número de ítems resueltos y el tiempo de latencia en las respuestas no son medidas equivalentes de la habilidad de los sujetos y será necesario tomar información de ambas variables.

– Las omisiones tampoco pueden ser tratadas de la misma manera, en los tests de potencia puros, las omisiones marcan el punto de corte de aquellos ítems cuyas respuestas no conoce el sujeto y que han de ser puntuadas como incorrectas. En los tests de velocidad puros, estas omisiones pueden ser tratadas como omisiones aleatorias no relacionadas con la habilidad medida por el test y deben ser eliminadas. Cuando los tests no son de velocidad ni de potencia puros, la interpretación correcta de estas omisiones es difícil de conocer.

– En los tests de rendimiento, aptitudes y habilidad, los sujetos tienden a responder bajo el principio de maximizar su puntuación para lo que han de utilizar distintas estrategias. Esto demuestra que es necesario tener en cuenta varias cosas, el tiempo de respuesta, el número de ítems que se han acertado y las diferentes estrategias a seguir para maximizar la puntuación. Ahora bien, para complicar un poco más la cosa, la elección de una estrategia u otra es posible que dependa no sólo del nivel de habilidad medida por el test, sino de determinados factores de personalidad que si no se tienen en cuenta en el modelo pueden ocasionar sesgos entre sujetos que hayan utilizado estrategias diferentes.

Todos estos problemas ponen de manifiesto la dificultad de desarrollar modelos para aquellos tests que no son ni de potencia ni de velocidad puros; sin embargo, es necesario investigar en este sentido y analizar los problemas que comportan.

Dentro de este apartado se recogen tres tipos de modelos, el *Modelo logístico para tests de tiempo límite* (Verhelst, Verstralen y Jansen), que tiene en cuenta tanto la velocidad como la potencia; este modelo asume una función logística derivada de una distribución gamma para tiempos de respuesta y una distribución generalizada para una respuesta latente dado un determinado tiempo

de respuesta. *Los modelos para tests de velocidad y tiempo límite* de Roskam, que asumen una función exponencial y se utilizan bastante en tests que miden destrezas psicomotoras y los *Modelos de respuesta múltiple-intento, único-ítem* (Multiple-Attemp, Single-Item, MASI) presentados por Spray para analizar datos procedentes de pruebas psicomotoras. Estos modelos se diferencian de los modelos clásicos de la TRI que suelen considerarse «único intento, múltiples ítems».

Estos últimos modelos, hacen fuertes asunciones acerca de la habilidad latente medida y de los repetidos intentos de los sujetos. En primer lugar, asumen el supuesto de *unidimensionalidad*, aunque en la práctica este supuesto se suele violar ya que muchas habilidades psicomotoras, que son las que se suelen medir con este tipo de tests, son probablemente multidimensionales. También suele violarse un segundo supuesto de los modelos TRI que es el de *independencia local* ya que en este tipo de pruebas, en las que hay varios intentos por parte de los sujetos, es muy difícil que no afecten a los intentos sucesivos los resultados obtenidos en los intentos previos. La violación de estos supuestos puede afectar significativamente a las estimaciones obtenidas. Será necesario investigar en este sentido para comprobar la robusted de las estimaciones frente a la violación de los supuestos.

El programa OPLM (One Parameter Logistic Model) de Verhelst, Glass y Verstralen (1995), permite la estimación de los parámetros.

Para el modelo de Roskam no hay software disponible, la estimación de los parámetros se hace por etapas, primero se comprueba si el tiempo de respuesta se ajusta al modelo de Rasch y, si eso es así, una vez estimados los parámetros para el tiempo de respuesta, se estiman los correspondientes al «poder mental».

Spray tiene elaborado un programa para la estimación de los parámetros de sus modelos MASI-IRT, no está comercializado pero puede adquirirse escribiendo al autor

Modelos para múltiples habilidades o componentes cognitivos:

Hemos ido viendo cómo a lo largo del tiempo los modelos iniciales de la TRI se han ido adaptando a las nuevas necesidades de medición en función de los problemas que iban surgiendo: aparición de nuevos formatos de ítems por ejemplo; pero el hecho es que algunos de estos nuevos formatos requieren que los sujetos utilicen más de una destreza: solución de problemas, razonamiento crítico, organización, escritura, etc. lo que lleva a pensar en la necesidad de que se desarrollen modelos multidimensionales dentro del marco de la TRI para poder representar adecuadamente la actuación de los sujetos en el test.

Lord y Novick (1968) y Samejima (1974) fueron los pioneros en este tipo de modelos y a partir de sus trabajos se han ido desarrollando otros muchos, a veces como extensiones de los modelos unidimensionales existentes. Mayden (1996) hace una clasificación de los distintos modelos en función del tipo de respuesta, el tipo de modelo (compensatorio o no compensatorio) y el tipo de función.

Algunos ejemplos a destacar serían: el *Modelo multidimensional de ojiva normal* (Mc Donald) que es una extensión del modelo unidimensional desarrollado por Lord (1952) y el *Linear logistic multidimensional model,* para datos dicotómicos (Reckase) desarrollado para extensiones multidimensionales de los modelos logísticos unidimensionales de Lord y Birnbaum. Fischer y Selinger proporcionan otros modelos que son extensiones multidimensionales de los modelos de Rasch para ítems dicotómicos y politómicos: *Modelos multidimensionales logísticos- lineares para la medida del cambio.* Las principales diferencias entre los dos

primeros modelos, el de Mc Donald y el de Reckase, y los de Fischer y Selinger son que estos últimos: a) se ciñen a los modelos de Rasch, b) están diseñados para medir el cambio, por lo que es necesario obtener datos de los mismos sujetos en dos ocasiones distintas, c) dan cuenta detallada de los procesos cognitivos que se ponen en marcha al responder a los ítems de un test.

Dentro de los modelos para ítems politómicos con categorías de respuesta ordenada merecen destacar el *Modelo logit multinomial de coeficientes aleatorios* (Adams, Wilson y Wang) que es una generalización de una serie de modelos logísticos de un parámetro y el *Modelo multidimensional de respuesta graduada* (Muraki y Carlson). Para categorías de respuesta no ordenadas Takane y Leeuw presentaron en 1987 un modelo. Kelderman (1996) proporciona un *Modelo multidimensional para el análisis del crédito parcial*

Además de estos modelos merecen citarse el *Modelo de respuesta al ítem multidimensional log-lineal* para ítems puntuados politómicamente (Kelderman) cuyo atractivo reside en la flexibilidad que proporcionan los modelos loglineales con sus procedimientos de estimación de parámetros bien desarrollados y *Los Modelos de respuesta con predictores manifiestos* (Zwinderman) que, aunque no son en si mismos modelos multidimensionales de TRI, permiten la estimación de una puntuación de habilidad unidimensional a partir de un conjunto de variables independientes. Esta puntuación será una puntuación ponderada a partir de esas variables independientes. El resultado es una especie de análisis de regresión logística que está siendo muy utilizado en el análisis de datos psicológicos y educativos.

Aunque se ha hecho un esfuerzo enorme en el desarrollo de estos modelos, como lo demuestran los numerosos trabajos que han ido apareciendo algunos de los cuales se han recogido en el número monográfico de la revista *Applied Psychological Measurement* (1996, Vol. 20, Nº 4), quedan todavía muchos temas por analizar.

También son importantes los modelos denominados de componentes cognitivos. Sabemos que el escribir los ítems de los tests ha sido una empresa bastante pragmática. La redacción de los ítems debía ajustarse a una serie de especificaciones en relación al contenido. A partir del análisis de ítems se comprueba que algunos de ellos proporcionan información útil, pero otros han de eliminarse. En algún momento es necesario hacer un estudio para averiguar qué es lo que miden los ítems del test. Este proceso se repite una y otra vez por los constructores de tests, pero sería bueno preguntarse acerca de ¿qué es lo que hace que un ítem sea difícil? ¿es posible que se construyan sistemáticamente ítems que impliquen la incorporación de distintas componentes o destrezas cognitivas? Estas son las preguntas a las que Fischer y Embretson han tratado de responder. Fischer, por ejemplo, introduce un grupo de modelos: *Modelos de Rasch unidimensionales logístico lineales*, en los que el parámetro de dificultad de los ítems se descompone, a lo largo del análisis, en una serie de parámetros para las destrezas cognitivas o componentes (parámetros básicos) necesarias para resolver el ítem. Estos modelos permiten un análisis más cuidadoso del material de evaluación y proporcionan las bases para una aproximación más sistemática al desarrollo de los tests, donde la validez de cada ítem forme parte del proceso de construcción de los tests y no algo que se investigue una vez construido el mismo.

Embretson con sus *Modelos de respuesta multicomponentes*, intenta proporcionar modelos multidimensionales que den cuenta de la actuación de los sujetos en tareas complejas, estimando la habilidad de los mismos en cada una de las subtareas necesarias para completar la tarea más compleja.

El programa NOHARM (Fraser,1988) permite el ajuste de los modelos de ojiva normal, tanto unidimensionales como multidimensionales, de la TRI. La estimación de los parámetros en los modelos multidimensionales logísticos de Reckase se puede hacer mediante el programa MAXLOG (McKinley y Reckase, 1983) y MIRTE (Carlson, 1987). McKinley (1987) desarrolló también otro procedimiento basado en el método de máxima verosimilitud marginal: MULTIDIM.

Para los modelos multidimensionales loglineares de Kelderman hay una serie de programas que pueden usarse: LOGIMO (Kelderman y Steen, 1993); OPLM (Verhelst et al.. 1993). Para los modelos de Embretson se puede utilizar el programa MULTICOMP (Embretson, 1984), COLORA (Maris, 1993).

La importancia de los trabajos de Fischer y Embretson estriba en que permiten abordar en un futuro una nueva forma de construir y analizar los tests.

Modelos no paramétricos

Los orígenes de estos modelos se encuentran en los primeros trabajos de Guttman (1947, 1950) sobre los análisis del escalograma, en los que se asumía que las funciones de respuesta a los ítems tenían la forma de una curva en escalera en las que la relación entre la respuesta correcta a un ítem y la habilidad subyacente es una relación determinística; es decir, se asumía que por encima de un cierto umbral en la escala, los sujetos tienen una probabilidad igual a la unidad de responder incorrectamente a un ítem, y por debajo de dicho umbral una probabilidad igual a la unidad de responderlo correctamente. Debido a la rigidez de este modelo y al hecho de que en la práctica pocos datos mostrarían un claro ajuste, se pensó que sería más útil un modelo estocástico con una función de respuesta continua y se hicieron varios intentos para formular tales funciones a nivel no paramétrico. Esto implica que no se especifica la forma de la función característica de los ítems. Estos intentos culminaron con el trabajo de Mokken (1971) quien además de proporcionar un modelo no paramétrico, proporcionó la teoría estadística necesaria para probar si las propiedades formales del modelo presentado se daban en los datos empíricos. Sin embargo hasta principios de los 80, principalmente a partir de los trabajos de Holland (1981), Holland y Rosenbaum (1986), y Rosenbaum (1984, 1987), el tópico de TRI no paramétrica no se reintrodujo en la teoría psicométrica. La idea no era tanto proporcionar modelos no paramétricos como una alternativa a los modelos paramétricos, como el estudiar los supuestos mínimos que han de cumplir cualquier modelo, sea paramétrico o no paramétrico. Esto tuvo un gran interés no solo a nivel teórico sino a nivel práctico ya que para cada uno de los supuestos se especificaba claramente las consecuencias observables y se podían interpretar mejor los desajustes de los datos. Teniendo en cuenta que los modelos no paramétricos están basados en unos supuestos mínimos, se puede asumir que las funciones de respuesta que proporcionan están más próximas a las verdaderas funciones que las que proporcionan los modelos paramétricos.

El modelo de Mokken es una alternativa a los modelos probabilísticos para ítems dicotómicos y, según él mismo indica, pueden ser utilizados para ordenar a los sujetos con respecto a su puntuación total en un conjunto de ítems monótonos-homogéneos, de manera que, con independencia del error de medida, refleje el orden de esos sujetos sobre la propiedad medida por el conjunto de ítems (habilidad, actitud, capacidad, rendimiento, etc.). Molenaar generaliza la teoría a ítems politómicos y, finalmente Ramsay, discute una serie de técnicas para estimar las funciones de respuesta de ítems dicotómicos basadas en supuestos ordinales.

El programa MSP de (Molenaar y colaboradores, 1994) permite el análisis de datos del modelo de Mokken para ítems politómicos

En algunas situaciones, como son el escalamiento de actitudes, en el análisis de la conducta de los votantes, en estudios de mercado, etc. los ítems son difíciles de obtener o escasos, o el nivel de información acerca de su calidad no es lo suficientemente alto como para garantizar el uso de modelos paramétricos; en este caso, los investigadores tienden a evaluar las actitudes, habilidades y las dificultades de los ítems asociadas, utilizando niveles de medida ordinales en lugar de intervalos o de razón requeridos por los modelos paramétricos

Modelos para ítems no monotónicos

Los primeros modelos que se desarrollaron en la TRI hacían referencia a variables de conocimientos, habilidades, destrezas, etc. en las que el supuesto de monotonicidad tenía sentido. Sin embargo, en las ciencias sociales y de la conducta hay algún dominio de conductas en las que es improbable que se de este supuesto. Se trata del dominio de las actitudes, opiniones, creencias y valores. En general este tipo de variables se miden mediante una serie de instrumentos formados por un conjunto de declaraciones, frases o preguntas, a las que los sujetos deben contestar indicando su grado de acuerdo o desacuerdo. La experiencia ha demostrado que en lo que se refiere a estas variables los extremos se tocan. No es infrecuente encontrar que personas con actitudes opuestas están de acuerdo con los mismos ítems. Por supuesto las razones de este acuerdo son contrarias en los dos tipos de personas. Esto violaría el supuesto de monotonicidad.

Dentro de este grupo de modelos podemos destacar el de Andrich (1997) que se trata de un modelo unidimensional de despliegue y que parte del hecho de que las funciones de respuesta de las categorías no extremas (intermedias), aquellas que producen ambigüedad en los sujetos, en los ítems politómicos también son no monotónicas. El modelo es el resultado de refundir un modelo de 3 categorías de clasificación en dos categorías de respuesta (acuerdo/desacuerdo). Otro de los modelos es el PARELLA de Hoijtink (1997) basado en el análisis de paralelogramo de Coombs, pero especifica la función de respuesta directamente como una función de densidad de Cauchy. El programa PARELLA permite la estimación de los parámetros en estos modelos.

Modelos con supuestos especiales acerca de los procesos de respuesta

Dentro de este apartado se incluyen una serie de modelos que no tienen entre sí unas características comunes; sin embargo, todos ellos son importantes para el análisis de un tipo especial de datos. Se trata de modelos para distintos procesos de respuesta o distintas distribuciones de habilidad, dependencia condicional entre las respuestas, formatos de respuesta diseñados para permitir el conocimiento parcial en los ítems del test, etc.

Los *Modelos TRI multigrupo* presentados por Bock y Zimowski (1997) permiten unificar algunas de las aplicaciones de la TRI en situaciones en las que debe aplicarse un determinado modelo a varios grupos para estudios de DIF, cuando se quieren unir varios tests o equipararlos a una escala común, etc. En general, en situaciones en las que existen personas procedentes de varias poblaciones y que responden al mismo test o a tests que tienen ítems comunes y donde el objetivo es estimar los parámetros de los ítems y la distribución latente de la habilidad común de las personas en cada una de las poblaciones.

Con algunas excepciones los procedimientos multigrupo para el análisis de ítems

dicotómicos se han implementado en el programa BILOG- MG de Zimowski y colaboradores (1996) y para múltiples categorías en el programa PARSCALE de Muraki y Bock (1991).

Hay otro tipo de modelos incluídos en este apartado que se denominan *Modelos logísticos mixtos*, presentados por Rost (1997) y que asumen que los datos observados no proceden de una población homogénea de sujetos sino de una mezcla de datos provenientes de dos o más poblaciones latentes. El objetivo de estos modelos es separar los datos en subpoblaciones homogéneas y estimar en cada una de ellas los parámetros correspondientes. Una de las características de este grupo de modelos es que las subpoblaciones no están definidas por variables manifiestas sino por variables latentes. Se trata de modelos que son mezcla de modelos TRI y un modelo de clase latente. Hasta la fecha se conocen trabajos realizados con el modelo de Rasch, denominándose *Modelo mixto de Rasch* a la generalización del modelo ordinario a un modelo de distribución mixta. Se ha trabajado tanto con datos dicotómicos como con datos politómicos (van Davier y Rost, 1995; Rost,1997 y Rost y van Davier, 1993, 1995). Estos modelos pueden ser analizados mediante el programa de Windows: WIN-MIRA (Mixed Rasch model de van Davier, 1995).

Uno de los supuestos fundamentales de la TRI era el supuesto de independencia local. Desafortunadamente este supuesto se viola, al menos en algún grado, en muchos de los tests que se utilizan. Por ejemplo cuando alrededor de un estímulo o pasaje común se organizan varios ítems. En este caso las respuestas a los ítems puede que no sean independientes unas de otras y se viole el supuesto de independencia local. Esto complica la estimación de los parámetros y la estimación de la habilidad puede resultar poco fiable. Una solución a este problema es desarrollar modelos TRI que no impliquen este supuesto. A estos modelos se les ha llamado; *Modelos localmente dependientes: TRI conjuntiva* . Estos modelos, desarrollados por Jannarone y sus colegas (1997), tanto como modelos paramétricos como no paramétricos, se diferencian de otros modelos TRI en que utilizan estadísticos suficientes no aditivos para los ítems y para las personas. La utilidad de estos modelos se ha puesto de manifiesto en algunas situaciones ya que, por ejemplo, permiten medir el aprendizaje durante el proceso de evaluación mientras que con los modelos convencionales no se puede hacer. Medir, por ejemplo, cómo responden los sujetos al refuerzo durante entrenamientos interactivos. En evaluaciones computarizadas, mediante sesiones interactivas, en las que se puede evaluar tanto el rendimiento como el tiempo de respuesta. En aplicaciones con tests adaptados a los sujetos, en los CAT. Como se puede deducir, son todas aplicaciones en las que se viola el supuesto de independencia local.

Finalmente cabe incluir en este apartado otro tipo de modelos para tests cuyo formato de respuesta permita el conocimiento parcial por parte de los sujetos y paliar así una de las críticas que se hacen a los tests de elección múltiple cuando se puntúan de forma dicotómica que es que no permiten la evaluación del conocimiento parcial. Estos modelos parten de que en las respuestas incorrectas de los tests de elección múltiple hay contenido un cierto conocimiento parcial, algunas respuestas incorrectas reflejan más conocimiento parcial que otras y el nivel de habilidad puede ser estimado más adecuadamente si se utiliza esta información parcial. Hutchinson y sus colaboradores (1997) describen una familia de modelos: *Mismatch models for Test formats that permit parcial information to be shown*, para estimar la habilidad de los sujetos y obtener además información acerca del funcionamiento cognitivo de los mismos en los

ítems del test. Estos modelos son nuevos y merecen ser investigados. También merece la pena investigar las diferencias entre las soluciones propuestas por Hutchinson para solucionar el problema del conocimiento parcial en los tests de elección múltiple y los otros modelos que hemos presentado para ítems politómicos.

Referencias

Adams, R.J. y Khoo, S.T. (1992). *QUEST: The interactive test analysis system.* Melbourne, Victoria: Australian Council for Educational Research.

Andersen, E.B. (1997). The rating scale model. En W.J. van der Linden y R.K. Hambleton (Eds.),*Handbook of Modern Item Response Theory*, 67-86. New York: Springer-Verlag.

Andrich, D. (1978) A rating formulation for ordered response categories. *Psychometrika* 43, 561-573.

Andrich, D. (1997). An hiperbolic cosine IRT model for unfolding direct responses of persons to items. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 399-414. New York: Springer-Verlag.

Barbero, I. y Prieto, P. (1995). Efectos de la violación de los supuestos del modelo de Rasch sobre la robusted de las estimaciones.*Psicotema*, 7, 2, 419-426.

Barbero, I. y Navas, M.J. (1995). *Creación de un sistema computerizado de evaluación de la capacidad matemática.* Centro de Investigación,Documentación y Evaluación (CIDE), Madrid.

Barbero, I. (1996). Los bancos de ítems. En J. Muñiz (Coord.), *Psicometría*, 139-170, Madrid: Universitas.

Barbero, I. y Prieto, P. (1997). Evaluación del rendimiento en Ciencias de los niños y niñas de 13 años de las distintas comunidades autónomas: Impacto o sesgo.*Psicothema*, 9 (2), 323-332.

Birnbaum, A.(1968). Some latent trait models and their use in inferring an examinee's ability. En F.M. Lord y M.R. Novick (Eds.), *Statistical theories of mental tests scores*, 397-479, Reading, MA: Addison-Wesley.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29 -51.

Bock, R.D. (1997). The nominal categories model. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 33-49. New York: Springer-Verlag.

Bock, R.D. y Zimowski, M.F. (1997). Multiple group IRT. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 433-448. New-York: Springer-Verlag.

Carlson, J.E. (1987). *MIRTE: Multidimensional item response theory estimation: A computer program* (Research report ONR 87-2) Iowa City, IA: American College Testing.

Cuesta, M. y Muñiz, J. (1994). Utilización de los modelos unidimensionales de teoría de respuesta a los ítems con datos multifactoriales. *Psicothema,* 6 (2), 283-296.

Cuesta, M. y Muñiz, J. (1995). Efectos de la multidimensionalidad en la estimación de parámetros desde modelos unidimensionales de teoría de respuesta a los ítems. *Psicológica,* 16 (1), 65-86.

Cuesta. M. (1996). Unidimensionalidad. En J. Muñiz (Coord.) *Psicometría*, 239-292, Madrid: Universitas.

Embretson (1984) MULTICOM: A general multicomponent latent trait model for response processes. *Psychometrica,* 49, 175-186.

Embretson, S.E. (1997). Multicomponent response models. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 305-321. New York: Springer-Verlag.

Ferguson, G.A. (1942). Item selection by the constant process. *Psychometrika,*7, 19-29.

Ferrando, P.J. (1996). Relaciones entre el análisis factorial y la teoría de respuesta a los ítems. En J. Muñiz (Coord.), *Psicometría,* 555-612. Madrid: Universitas.

Fidalgo, A.M. (1996). Funcionamiento diferencial de los ítems. En J. Muñiz (Coord.), *Psicometría,* 371 156. Madrid: Universitas.

Fischer, G.H. y Seliger; E. (1997). Multidimensional linear logistic models for change. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 323-346. New York: Springer-Verlag.

Fischer, G,H, (1997). Unidimensional linear logistic Rasch models. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 225-243. New York: Springer- Verlag.

Fraser, C. (1986). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, Australia: The University of New England.

Gómez, J. e Hidalgo, M.D. (1997). *A comparison of two procedures of ability purification on the detection of differential item functioning using multinomial logistic regression.* Comunicación presentada en el 10th European Meeting of the Psychometric Society, Santiago de Compostela.

Gómez, J. y Navas, M.J. (1996) Detección del funcionamiento diferencial de los ítems mediante regresión logística: Purificación paso a paso de la habilidad. *Psicológica*, 17, 397-411.

Guttman, L. (1947). The Cornell technique for scale and intensity analysis. *Educational and Psychological Measurement*, 7, 247-280.

Guttman, L. (1950). Relation of scalogram analysis to other techniques. En S.A. Stouffer; L. Guttman; E.A. Suchman; P.F. Lazarsfeld; S.A. Star y J.A. Clausen (Eds.),*Measurement and Prediction: Studies in Social Psychology in World War II* (Vol. 4). Princenton. NJ: Princenton University Press.

Haley, D.C.(1952). *Estimation of the dosage mortality relationship when the dose is subject to error.* (Technical Report No. 15). Palo Alto, CA: Applied Mathematics and Statistics Laboratory, Stanford University.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. En R.L. Linn (Ed.), *Educational Measurement*, 147 - 200, New York: Macmillan.

Hidalgo, M.D. y López-Pina, J.A. (1997). *Detección del DIF en ítems politómicos e igualación iterativa: Comparación entre las medidas de área de Raju y el estadístico de Lord.* Comunicación presentada en el V Congreso de Metodología de las Ciencias Humanas y Sociales. Sevilla.

Hoijtink, H. (1990). A latent trait model for dichotomous choice data. *Psychometrika*, 55, 641-656.

Hoijtink, J.; Molenaar, I.W. y Prost, W.J. (1994). *PARELLA: User's manual.* Groningen, the Netherlands: iec ProGAMMA.

Hoijtink, H. (1997). PARELLA: An IRT model for parallelogram analysis. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 415-429. New York: Springer- Verlag.

Holland, P.W. (1981). When are item response models consistent with observed data? *Psychometrika*, 46, 79-92.

Holland, P.W. y Rosembaum, P.R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics*, 14, 1.523-1.543.

Hutchinson, T.P. (1997). Mismatch models for test formats that permit partial information to be shown .En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 481-494. New York: Springer-Verlag.

Jannarone, R.J. (1991). Conjunctive measurement theory: Cognitive research prospects. En M. Wilson (Ed.), *Objetive Measurement: Theory into practice*. Norwood, NJ: Ablex, 211-236.

Jannarone, R.J. (1997). Models for locally dependent responses: Conjunctive item response theory. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 465 -479. New York: Springer-Verlag.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika*, 49, 223-245.

Kelderman, H.y Steen, R. (1993). *LOGIMO: Loglinear item response modeling* (computer manual). Groningen, the Netherlands: iec ProGAMMA.

Kelderman, H. (1997). Loglinear multidimensional item response models for polytomously scored items. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 287-304. New York: Springer-Verlag.

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

López-Pina, J.A. (1995). *Teoría de respuesta a los ítems: Fundamentos.* Barcelona: PPU.

López-Pina, J.A. e Hidalgo, M.D. (1996). Bondad de ajuste y teoría de respuesta a los items. En J. Muñiz (Coord.). *Psicometría*, 643-704. Madrid: Universitas.

Lord, F.M. (1952). A theory of mental test scores. *Psychometric Monograph, No. 7.*

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

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika,* 58, 445-469.

Martínez-Cardeñoso, J., Cuesta, M. y Muñiz, J. (1996). Dimensionalidad y función de información de los tests. *Psicothema,* 8 (1), 215-220.

Masters. G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika,* 47, 149-174.

Masters, G.N. y Wright, B.D. (1984). The essential process in a family of measurement models. *Psychometrika,* 49, 529-544.

Masters, G.N. y Wright, B.D. (1997). The partial credit model. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory,*101-121. New York: Springer-Verlag.

Maydeu, A. (1996). Modelos multidimensionales de teoría de respuesta a los ítems. En J. Muñiz (Coord.). *Psicometría,* 811-868. Madrid: Universitas.

Mc Donald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs,* No. 15.

Mc Donals, R.P. (1985). Unidimensional and multidimensional models for item response theory. En D.J. Weiss (Ed.), *Proceedings of the 1982 item response theory and computarized adaptative testing conference,* 127-148. Minneapolis, MN: University of Minnesota.

Mc Donald, R.P. (1997). Normal ogive multidimensional model. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory,* 257-269. New York: Springer-Verlag.

McKinley, R.L. y Reckase, M.D. (1983) MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods and Instrumentation,* 15, 389-390.

McKinley, R. L.(1987). *User's guide to MULTIDIM.* Princenton, NJ: Educational Testing Service.

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

Mislevy, R.J. y Bock, R.D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models.* (computer program). Chicago, IL: Scientific Software International.

Mokken, R.J. (1971). *A theory and procedure of scale analysis, with applications in political research,* New York/Berlin: Walter de Gruyter-Mouton.

Mokken, R.J. (1997). Nonparametric models for dichotomous responses. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory,* 351-367. New York: Springer-Verlag.

Molenaar, I.W.; Debets, P.; Sijtsma, K. Y Hemker, B.T. (1994). *MSP, a program for Mokken scale analysis for polytomous items, version 3.0. (user's manual).* Groningen, the Netherlands: iec ProGAMMA.

Molenaar, I.W. Nonparametric models for polytomous responses. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory,* 369-380. New York: Springer-Verlag.

Molina, G. (1997). *Los bancos de ítems en el desarrollo de tests. Aspectos psicométricos y análisis de un sistema para su desarrollo y gestión informatizada.* Tesis doctoral. Universidad de Valencia.

Mosier, C.I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review,* 47, 355-366.

Mosier, C.I. (1941). Psychophysics and mental test theory II: The constant process. *Psychological Review,* 48, 551-560.

Muñiz, J., Rogers, J. y Swaminathan, H. (1989) Robustez de las estimaciones del modelo de Rasch en presencia de aciertos al azar y discriminación variable de los ítems. *Anuario de Psicología,* 4 (3), 83-97.

Muñiz, J. (1990). *Teoría de respuesta a los ítemes. Un nuevo enfoque en la evolución psicológica y educativa.* 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. (1996). *Psicometría.* Madrid: Universitas.

Muñiz, J. (1997). Introducción a la teoría de respuesta a los ítems. Madrid. Pirámide.

Muraki, E. y Bock, R.D. (1991). *PARSCALE: Parametric Scaling of Rating Data.* Chicago: Scientific Software International

Muraki, E. (1992) A generalized partial credit model: Application of an EM algorithm *Applied Psychological Measurement,* 16, 159-176.

Muraki, E. (1997). A generalized partial credit model. En W.J. van der Linden y R.K. Ham-

bleton (Eds.), *Handbook of Modern Item Response Theory*, 153-164. New York: Springer-Verlag.

Navas, M. J. (1993). *Aplicación de la teoría de respuesta al ítem al campo de la medida: Creación de un banco de ítems para evaluar la capacidad matemática.* Tesis doctoral. UNED. Madrid.

Navas, M.J. (1994). Teoría clásica de los tests versus teoría de respuesta al ítem. *Psicológica,*15, 175-208.

Navas, M.J. (1996). Equiparación de puntuaciones. En J. Muñiz (Coord.) *Psicometría,* 293-370. Madrid: Universitas.

Olea, J. y Ponsoda, V. (1996). Tests adaptativos informatizados. En J. Muñiz (Coord.).*Psicometría*, 729-784. Madrid: Universitas.

Ponsoda, V., Olea, J. y Revuelta, J. (1994). ADTEST. A computer- adaptative test based on the maximum information principle. *Educational and Psychological Measurement,* 54, 3, 680-686.

Prieto, P. y Barbero, I. (1996). Detección del funcionamiento diferencial de los ítems mediante el análisis de residuales: Una aplicación de la TRI. *Psicothema,*8 (1), 173-180.

Prieto, P., Barbero, I. y San Luís, C. (1997). Identification of nonuniform DIF: A comparison of Mantel-Haenszel and URT analysis procedures. *Educational and Psychological Measurement,* 57 (4), 559-568.

Prieto, G. y Delgado, A. (1996). Construcción de ítems. En J. Muñiz (Coord.) *Psicometría,* 105-138. Madrid: Universitas.

Ramsay, J.O. (1992). *TESTGRAF: A program for the graphical analysis of multiple choice test and questionnaire data.* (Technical Report) Montreal, Quebec: McGill University.

Ramsay, J.O. y Wang, X. (1993). *Hybrid IRT models.* Paper presented at the meeting of the Psychometric Society, Berkeley, CA.

Ramsay, J.O. (1997). A functional approach to modeling test data. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 381-397. New York: Springer-Verlag.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 271-286. New York: Springer-Verlag.

Renom, J. (1993). *Tests adaptativos computerizados: Fundamentos y aplicaciones.* Barcelona: PPU.

Renom, J. (1998). *Tratamiento informatizado de datos.* Barcelona: Masson.

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

Rosenbaum, P.R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrica,* 49, 425-435.

Rosenbaum, P.R. (1987). Comparing item characteristic curves. *Psychometrika,* 52, 217-233.

Rost, J. y van Davier, M. (1993). Measuring different traits in different populations with the same items. En R. Steyer; K.F. Wender y K.F. Widaman (Eds.), *Psychometric Methodology, Proceedings of the 7th European Meeting of the Psychometric Society,* 446-450. Stuttgart/ New York: Gustav Fischer Verlag.

Rost, J. y van Davier, M. (1995). Mixture distribution Rasch models. En G. Fischer e I. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications,* 257- 268. New York: Springer-Verlag.

Rost, J. y Davier, M. van. (1992). *MIRA: A PC program for the mixed Rasch model. (User manual).* Kiel, Germany: Institute of Science Education (IPN).

Rost, J. (1997). Logistic mixture models. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 449-463. New York: Springer-Verlag.

Roskam, E.E. (1997). Models for speed and time limit tests. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 187-208. New York: Springer-Verlag.

Samejima, F. (1974). Normal ogive model for the continuous response level in the multidimensional latent space. *Psychometrika,* 39, 111-121.

Samejima, F. (1979). *A new family of models for the multiple choice item* (Research Report No 79-4). Knoxville, TN: Department of Psychology, University of Tennessee.

Samejima, F. (1997). Graded response model. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*..85-100. New York: Springer-Verlag.

San Luis, C., Prieto, P., Sánchez-Bruno, A. y Barbero, I. (1995). GENESTE: Un programa de control para TRI. *Psicológica*, 16 (2), 297-304.

Spray, J. A. (1997). Multiple- attempt, single- item response models. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 209-220. New York: Springer-Verlag.

Thissen, D. y Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, 49, 501-519.

Thissen, D. (1991). *MULTILOG User's guide, version 6.* Chicago Scientific Software International.

Thissen, D. y Steinberg, L.(1997). A response model for multiple choice items. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 51- 65. New York: Springer-Verlag.

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review*, 34, 278- 286.

Tutz, G. (1997). Sequential models for ordered responses. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*,139-152. New York: Springer-Verlag.

Van Davier, M.V. y Rost, J. (1995). *WIN-MIRA: A program system for analysis with the Rasch model, with the latent class analysis and with the mixed Rasch model.* Kiel: Institute for Science Education (IPN), distributed by Icc ProGAMMA, Groningen.

Van der Linden y Hambleton (1997). Handbook of modern item response theory. New York: *Springer-Verlag.*

Verhelst, N.D.; Glas, C.A.W. y de Vries, H.H. (1997). A steps model to analyze partial credit. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory.*, 123-138. New York: Springer-Verlag.

Verhelst, N.D.; Verstralen, H.H.F.M. y Jansen, M.G.H. (1997). A logistic model for time-limit tests. En W.J. van der Linden y R.K. Hambleton(Eds.), *Handbook of Modern Item Response Theory*, 169-185. New York: Springer-Verlag.

Verhelst, N.D.; Glas, C.A.W. y Verstralen, H.H.F.M. (1995). *OPLM: one parameter logistic model. Computer program and manual.* Arnhem: Cito.

Wingersky, M.S.; barton, M.A. y Lord, F.M. (1982). *LOGIST User's Guide.* Princenton, NJ: Educational Testing Service.

Wright, B.D. y Masters, G.N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B.D. y Linacre, J.M. (1992). *BIGSTEPS Rasch analysis computer program.* Chicago: MESA Press

Zimowski, M.F.; Muraki, E.; Mislevy, R.J. y Bock, R.D. (1996). *BILOG-MG: Multiple-Group IRT analysis and test maintenance for binary items.* Chicago: IL: Scientific Software International

Zwinderman, A. H. (1997). Response models with manifest predictors. En W.J. van der Linden y R.K. Hambleton(Eds.), *Handbook of Modern Item Response Theory.*,245-256. New York: Springer -Verlag.

Aceptado el 11 de mayo de 1998