Psykhe
ISSN: 0717-0297
psykhe@uc.cl
Pontificia Universidad Católica de Chile
Chile

Asún Inostroza, Rodrigo; Zúñiga Rivas, Claudia
Ventajas de los Modelos Politómicos de Teoría de Respuesta al Ítem en la Medición de Actitudes Sociales. El Análisis de un Caso
Psykhe, vol. 17, núm. 2, noviembre, 2008, pp. 103-115
Pontificia Universidad Católica de Chile
Santiago, Chile

PSYKHE
2008, Vol.17, Nº 2, 103-115

ISSN 0717-0297

Ventajas de los Modelos Politómicos de Teoría de Respuesta al Ítem en la Medición de Actitudes Sociales. El Análisis de un Caso

Advantages of Polytomous Models of Item Response Theory in Measuring Social Attitudes. A Case Study

Rodrigo Asún y Claudia Zúñiga
Universidad de Chile

A pesar de sus ventajas en la investigación psicométrica, la Teoría de Respuesta al Ítem (TRI) no ha logrado imponerse en la práctica cotidiana de medición de constructos psicológicos o actitudes sociales. En esta investigación se muestra la utilidad de trabajar con modelos de TRI politómicos a través de su comparación con la Teoría Clásica, al estudiar el comportamiento de una escala de intolerancia. Se concluye que las ventajas de los modelos politómicos de la TRI no se encuentran en la forma en que escalan a los individuos ni en la estimación de los parámetros de los ítems, sino en la obtención de mayor información respecto al funcionamiento del instrumento, que podría ser utilizado para su mejoramiento futuro.

Palabras clave: *medición de actitudes, Teoría de Respuesta al Item, psicometría, modelos politómicos, intolerancia*.

Despite its advantages in psychometric research, Item Response Theory (IRT) has not been regularly used in the measurement of psychological constructs or social attitudes. This study used a scale of intolerance to demonstrate the usefulness of working with polytomous models of IRT in comparison with the Classical Test Theory to study attitudes. It is concluded that the advantages of the polytomous models of IRT are found not in the form in which the people are scaled or in the estimation of item parameters, but in obtaining better information of the psychometric proprieties of an instrument, information that can be used to improve the instrument.

Keywords: *attitude measurement, Item Response Theory, pyschometrics, polytomous models, intolerance.*

Presentación

Desde que fuera formulada alrededor de los años 60, la Teoría de Respuesta al Ítem (TRI) ha tendido a imponerse en la investigación psicométrica, aunque no en la práctica cotidiana de medición de constructos sociales y psicológicos. A continuación intentaremos explicar esta paradoja.

En una primera mirada, la TRI dispone de numerosas ventajas sobre la Teoría Clásica de los Test (TCT). Una presentación clásica de estas ventajas la realizó McKinley (1989). Mientras la TCT incorpora una serie de supuestos no posibles de poner a prueba, la TRI ofrece la posibilidad de contar con modelos cuyos supuestos sí pueden ser juzgados empíricamente, con lo que podemos formarnos un juicio respecto de la validez de las estimaciones obtenidas.

Además, mientras en la TCT las estimaciones de parámetros, tanto de los individuos como de las afirmaciones, dependen del grupo al que se ha aplicado el instrumento, en la TRI las estimaciones son invariantes, es decir, no dependen del grupo concreto al que se ha aplicado el test. Evidentemente, esto facilita la evaluación de las características de un instrumento y la comparación de los resultados.

Más importante aún, el cálculo de la consistencia interna (como indicador de confiabilidad) se realiza en la TCT para el conjunto de los individuos y el instrumento, con lo que obtenemos un índice global que suponemos propiedad del instrumento completo. En cambio, la TRI permite disponer de medidas de error distintas para cada individuo y/o nivel de habilidad, con lo que podemos saber cuan precisamente

Rodrigo Asún Inostroza, Departamento de Sociología, Universidad de Chile, Santiago, Chile.
Claudia Zúñiga Rivas, Departamento de Psicología, Universidad de Chile, Santiago, Chile.
La correspondencia relativa a este artículo debe ser dirigida a Rodrigo Asún Inostroza, Departamento de Sociología, Universidad de Chile, Ignacio Carrera Pinto 1045, Ñuñoa, Santiago, Chile. E-mail: rasun@uchile.cl

hemos medido a cada persona. En otras palabras, en la TRI se supone que el instrumento puede ser efectivo para medir a personas con cierto nivel de habilidades, pero deficiente para medir a personas con niveles diferentes.

Finalmente, en la TRI las estimaciones, tanto de las características de los ítems como de los individuos, se obtienen en la misma escala de medición, lo que facilita su comparación, mientras que en la TCT las estimaciones se obtienen en diversas escalas.

Producto de estas ventajas, desde hace algún tiempo la TRI es la teoría hegemónica en la investigación en psicometría (Santisteban & Alvarado, 2001), pues ella facilita el diseño de instrumentos más complejos que la TCT. Es así como diversos modelos de la TRI se han utilizado con mucho éxito en el diseño de test de rendimiento educativo de aplicación recurrente, en la generación de test adaptativos informatizados (TAIs), en la investigación sobre funcionamiento diferencial de los ítems (DIF) y en la construcción de bancos de ítems, entre otros temas (Muñiz, 1997). Además, cotidianamente se publican investigaciones que ponen a prueba sus supuestos bajo las más diferentes condiciones, proponen nuevos modelos o diseñan algoritmos alternativos de estimación de sus parámetros y programas informáticos para aplicarlos.

Sin embargo, a pesar de este evidente éxito a nivel académico, la TRI no ha logrado imponerse en todos los campos. En el plano de la medición aplicada aún se utiliza frecuentemente la tradicional TCT (Santisteban & Alvarado, 2001). No cabe duda que los altos costos que implica la utilización de la TRI, tanto en términos del tamaño de muestra necesario para calibrar sus parámetros como en los conocimientos matemáticos que exige su comprensión, entre otros problemas, han limitado su uso a situaciones muy específicas. Algo similar ha ocurrido en el campo de la medición de constructos no cognitivos (actitudes o personalidad), en la que se utilizan frecuentemente formatos de respuesta politómica.

Los formatos de respuesta politómica son aquellos en que se puede responder a cada afirmación en tres o más alternativas de respuesta, mientras que los formatos dicotómicos son aquellos en que solo se presentan dos alternativas de respuesta. Usualmente los formatos dicotómicos son utilizados en la medición del rendimiento o habilidades, pues lo que interesa es distinguir la capacidad de las personas de acertar la respuesta correcta (si hubiera más de una respuesta incorrecta, como en los ítems de selección múltiple, todas ellas se tienden a tratar indiferenciadamente como la misma respuesta incorrecta). En este sentido, uno de los principales parámetros que caracterizan a cada afirmación es su *grado de dificultad*, entendiéndose este como el nivel de conocimiento o habilidad que es necesario poseer para tener altas probabilidades de escoger la respuesta correcta.

Por su parte, los formatos politómicos tienden a ser utilizados en escalas de personalidad o actitudes, pues no se supone que existan respuestas correctas o incorrectas, sino que la diferencia entre las alternativas de respuesta es la intensidad con la que se debe poseer el constructo para responder cada una de ellas. En este caso, no existe propiamente un grado de dificultad del ítem sino, más bien, un *parámetro de posición o localización* de cada alternativa de respuesta, que indica la intensidad con que se debe poseer la característica medida para tener una alta probabilidad de responder cada alternativa.

Si bien los modelos de la TRI diseñados para trabajar con ítems dicotómicos tienen una importante difusión en investigación, no ocurre lo mismo con los propuestos para tratar con afirmaciones politómicas. Una solución usual para no tener que utilizar estos modelos ha sido agrupar las respuestas en dos categorías y luego aplicar los más conocidos y simples modelos dicotómicos, pero dicha solución puede no compensar la pérdida de información que supone el proceso.

Los aún escasos estudios que han aplicado modelos politómicos a la medición de actitudes o personalidad se han interesado fundamentalmente en los siguientes aspectos:

1. La aplicación de modelos politómicos de la TRI a test diseñados a partir de la TCT y verificación de su ajuste. En esta área se han obtenidos resultados contradictorios: en tanto unos estudios han encontrado que los modelos politómicos más usuales ajustan bien a instrumentos que evalúan actitudes o personalidad (Gómez, Hidalgo & Tomás-Sábado, 2007; Gray-Litter, Williams & Hancok, 1997; Roberts & Laughlin, 1996; Rojas & Lozano, 2005), otras investigaciones (Chernyshenko, Stark, Chan, Drasgow & Williams, 2001) ponen en duda que estos modelos puedan ajustar a instrumentos no cognitivos.
2. La determinación del número óptimo de alternativas de respuesta, respecto de lo cual los estudios han mostrado que, en ocasiones, utilizar cuatro alternativas de respuesta o incluso tres puede ser lo adecuado, resultando poco relevante la

información aportada por las otras opciones (Gray-Litter et al, 1997; Hernández, Muñiz & García, 2000).

3. La demostración de las ventajas de los modelos politómicos para el diseño TAIs eficientes (Hol, Vorst & Mellenbergh, 2005; Lai, Cella, Chang, Bode & Heinemann, 2003; Van Rijn, Eggen, Hemker & Sanders, 2002), para la medición del DIF (Kim, Cohen, Alagoz & Kim, 2007) y para la igualación de puntuaciones (Lee, Kolen, Frisbie & Ankenmann, 2001).
4. La prueba de nuevos modelos politómicos no paramétricos (Stout, 2001) o mejor adaptados a las escalas Likert (Javaras & Ripley, 2007).

No obstante lo anterior, el cálculo de puntajes directos y la TCT siguen dominando ampliamente el trabajo práctico en este campo, a pesar de las ventajas de los modelos politómicos de la TRI (Embretson & Reise, 2000), por lo que nos preguntamos: ¿serán estas relevantes en la utilización habitual de una escala de actitud? En otras palabras, en esta investigación no nos interesaron las potenciales ventajas que tendría el uso de modelos politómicos de la TRI en la construcción de instrumentos complejos, sino en la aplicación cotidiana de un instrumento a una muestra de tamaño regular, con el fin de diagnosticar el nivel de una actitud en una población determinada.

Es importante considerar que cuando se aplica una escala de actitud a una muestra de personas, se busca conocer habitualmente: (a) el escalamiento de las personas, (b) el grado en que la muestra o subgrupos de ella poseen la actitud medida, (c) el grado de asociación entre la actitud y otras variables y (d) la calidad psicométrica del instrumento utilizado.

De antemano conocemos algunas de las ventajas de utilizar la TRI y que hemos resumido anteriormente: podremos estimar el error de medición para cada individuo y no considerarlo solo como una propiedad global del instrumento, se obtendrán estimaciones invariantes en la misma escala y se podrá poner a prueba el ajuste de los modelos a los datos. No obstante, nos interesa saber si la aplicación de modelos politómicos de la TRI supondrá ventajas adicionales que justifiquen su utilización en una situación "habitual", como la descrita anteriormente.

En términos formales, nuestra pregunta de investigación fue: ¿Qué ventajas ofrece la aplicación de modelos politómicos de la TRI a una escala de actitud sobre el escalamiento de los individuos, la determinación de los parámetros de los ítems, la medición de la calidad psicométrica del instrumento utilizado y el grado de asociación entre la actitud y variables adicionales?

Ya que lo que nos interesaba era conocer la respuesta a esta pregunta en situaciones lo más reales posibles, hemos estudiado un caso concreto: aplicamos una serie de modelos de la TRI a una escala de actitud diseñada en Chile para medir el nivel de intolerancia de las personas frente a conductas, culturas y grupos alternativos o minoritarios.

Para efectos del presente estudio, decidimos utilizar los siguientes modelos TRI politómicos: (a) Modelo de Respuesta Graduada o de Samejima (MRG), (b) Modelo de Respuesta Graduada Modificado o de Muraki (MRGM), (c) Modelo de Crédito Parcial (MCP), (d) Modelo de Crédito Parcial Generalizado (MCPG) y (e) Modelo de Respuesta Nominal (MRN).

Se eligieron estos modelos en función de que algunos ya habían sido aplicados a test no cognitivos con cierto éxito (especialmente el MRG) y sus supuestos resultaban plausibles en escalas tipo Likert (respuestas ordenadas en el caso de MRG, MCP y MCPG e iguales opciones para todos los ítems en el modelo MRGM). Por su parte, el MRN, si bien está pensado para ítems de respuestas nominales, justamente por ello podía servirnos para conocer la existencia de desorden en las opciones de respuesta.

También hemos incorporado dos modelos de TRI dicotómicos: (a) el Modelo Logístico de Un Parámetro o de Rasch (1P) y (b) el Modelo Logístico de Dos Parámetros (2P). Esto lo hemos hecho para comparar la eficacia de los modelos politómicos sobre los más simples modelos dicotómicos.

A continuación se describen brevemente las características de cada uno de los modelos seleccionados.

Modelo Logístico de Un Parámetro o de Rasch

Este modelo permite el estudio de test compuestos por ítems dicotómicos (Rasch, 1960). Supone que la probabilidad de acertar una pregunta (o, en el caso de ítems actitudinales, dar la respuesta que implica presencia del constructo medido) depende solamente del poder discriminador de los ítems (que es constante para todos ellos) y de la dificultad o localización de cada afirmación en el continuo actitudinal. El *poder discriminador* es la capacidad de cada ítem de separar a individuos que poseen niveles distintos del constructo medido, asignándoles punta-

jes distintos, de manera que un ítem con *bajo* poder discriminador posiblemente asignará puntuaciones muy similares a personas que tienen niveles efectivamente diferentes, mientras que otro con *alto* poder discriminador hará lo contrario. Este modelo tiene el supuesto que todos los ítems del instrumento tienen el mismo poder discriminador (supuesto quizá poco realista, pero que simplifica enormemente las estimaciones). Por su parte, la dificultad o localización de cada afirmación es el nivel de habilidad o actitud que debe tener una persona para tener una probabilidad de 0,5 de acertar el ítem o dar la respuesta que implica la presencia de la actitud medida. Por ello, si un ítem tiene un valor mayor en este parámetro, será más difícil de acertar o se requerirá poseer mayor monto de la actitud medida para dar la respuesta que indica poseer dicha actitud.

El modelo de Rasch posee una serie de propiedades matemáticas y simplicidad que hacen muy deseable su utilización y, a la vez, requiere un menor tamaño de muestra para su correcta estimación (Santisteban & Alvarado, 2001). Lamentablemente, su misma simplicidad explica que pocas veces las respuestas de los individuos se ajusten al modelo.

Modelo Logístico de Dos Parámetros

Con el fin de contar con un modelo de supuestos menos restrictivos, se formuló el Modelo 2P (Birnbaum, 1957; Lord, 1952). La única diferencia entre este modelo y el anterior es que permite ítems con distinto poder discriminador. Ello explica que este modelo tenga, en ocasiones, un mejor ajuste que el modelo de Rasch, pero el costo de esta flexibilidad es requerir un mayor tamaño de muestra para estimar sus parámetros.

Modelo de Respuesta Graduada o de Samejima

El MRG o de Samejima (1969) fue creado para analizar ítems politómicos ordinales. En él se presupone que las alternativas de respuesta se encuentran ordenadas desde una menor a una mayor cercanía a un objeto actitudinal.

El MRG o de Samejima es una versión politómica del modelo de dos parámetros. Un problema de este modelo es que requiere la estimación de una gran cantidad de parámetros, pues se calcula un parámetro de localización distinto para cada transición entre alternativas de respuesta para cada ítem. Cada uno de estos parámetros expresa el grado de actitud en el cual se comienza a hacer más probable responder a una alternativa de respuesta, respecto de la alternativa de respuesta ordinal anterior.

Modelo de Respuesta Graduada Modificado de Muraki

El MRGM fue propuesto por Muraki (1990) como una forma simplificada del modelo de Samejima. Este modelo también tiene el supuesto que las categorías de respuesta se encuentran ordenadas. No obstante, a diferencia del modelo anterior, el de Muraki tiene el supuesto que todos los ítems tienen las mismas categorías de respuesta y que los parámetros de localización de las opciones de respuesta se encuentran separados por la misma distancia a lo largo de todos los ítems (por lo que se estima el mismo parámetro de localización para cada transición entre alternativas de respuesta para todos los ítems). Se puede decir que la diferencia subjetiva entre una categoría de respuesta y otra se supone constante para todas las afirmaciones, diferenciándose estas solo en la intensidad de actitud que es necesario poseer para marcar cada respuesta.

Una ventaja de este modelo más restrictivo es que requiere estimar menos parámetros, por lo que disminuyen los requerimientos de tamaño muestral para una correcta estimación.

Modelo de Crédito Parcial de Masters

El MCP es un modelo diseñado por Masters (1982) para ítems ordinales de *división por el total*. Los modelos de división por el total son aquellos en que la probabilidad de responder a una categoría determinada es calculada directamente. En este caso, la probabilidad que un individuo responda a una determinada categoría se obtiene dividiendo el numerador exponencial que corresponde a la categoría por la suma de los numeradores exponenciales de todas las categorías.

Modelo de Crédito Parcial Generalizado de Muraki

El MCPG de Muraki (1992) es una generalización del modelo anterior y, por tanto, está diseñado para ítems ordinales. La principal diferencia que tiene con el MCP es que se supone que la probabilidad de respuesta de un individuo a un ítem depende también de un *parámetro de discriminación*. Dicho parámetro representa la facilidad con que se puede pasar de una categoría a otra en cada ítem según el nivel de actitud que posean los individuos. Cuando

el parámetro de discriminación tiene un valor *alto*, indica que, dado un cierto nivel de actitud, es muy probable que las personas respondan a una alternativa de respuesta y no a las otras. Si este parámetro tiene un valor *bajo*, posiblemente las personas pueden utilizar distintas alternativas de respuesta, aunque posean la actitud o rasgo en el mismo grado.

Un problema de este modelo es que, al ser menos restrictivo que el MCP, se requiere estimar más parámetros.

Modelo de Respuesta Nominal de Bock

Este es un modelo propuesto por Bock (1972) para tratar ítems nominales, es decir, no se supone que exista orden en las opciones.

Este modelo exige el cálculo de muchos parámetros, por lo que requiere una muestra grande para realizar estimaciones precisas. Como contrapartida, tiende a ajustar frecuentemente a las respuestas de los individuos, porque impone pocas restricciones a los datos.

Objetivos

Los objetivos de este estudio fueron: (a) comparar el escalamiento de los individuos producido por la TCT y la TRI; (b) comparar la estimación de parámetros de los ítems que se obtienen en la TCT y la TRI; (c) determinar la calidad psicométrica del instrumento, estimada por la TRI y la TCT; y (d) determinar el grado de asociación entre la actitud y variables adicionales en la TRI y la TCT.

Método

Instrumento y Muestra

Se utilizó la base datos de la aplicación de una escala de actitudes para medir el nivel de intolerancia de las personas, denominada Escala de Intolerancia. Dicha escala fue diseñada por académicos de la Universidad de Chile el año 2001 y aplicada ese mismo año a 1.111 personas seleccionadas según un diseño muestral probabilístico polietápico (con cuatro etapas de selección aleatoria: comunas, manzanas, viviendas y personas) de habitantes mayores de 18 años de la Región Metropolitana. La muestra tuvo un promedio de edad de 39 años (DS = 16,23, rango de 18 a 92 años) y estuvo compuesta en un 47% por hombres. Respecto al nivel socioeconómico, se incluyó a todos los segmentos sociales desde el ABC1 hasta el E, con predominio de los sectores C3 (48% de la muestra).

El constructo a medir se definió como "la actitud de negar o restringir la posibilidad de expresar opiniones, valores y actuar conductas diferentes a las consideradas adecuadas y dar un trato discriminatorio a determinadas categorías sociales" (Aymerich, 2001, p. 1).

En términos de contenidos, se consideró que el concepto de intolerancia incluía: dogmatismo, autoritarismo, militarismo, sexismo, intolerancia religiosa, homofobia, clasismo, patriocentrismo, racismo e intolerancia hacia las minorías. A pesar de esta diversidad temática, se hipotetizó que el núcleo del constructo era potencialmente unidimensional.

La escala original estaba formada por 63 afirmaciones, a las que se agregaron algunas variables de identificación socio-demográfica. Frente a cada afirmación se presentaron seis alternativas de respuesta tipo Likert, en las que la respuesta 1 implicaba estar *muy de acuerdo* con la frase y la 6, *muy en desacuerdo*. Se incorporó además la posibilidad de responder *no sé* aunque, por no ser leída dicha alternativa a los individuos, solo representó el 3,7% de las respuestas.

Plan de Análisis

Se calcularon las puntuaciones directas y se utilizó la TCT para estimar la discriminación de los ítems y la consistencia interna del instrumento. Luego, se calibraron los modelos de la TRI dicotómicos y politómicos para determinar su ajuste a los datos y comparar sus resultados con la TCT.

Se siguieron las siguientes etapas en el análisis de la información.

Depuración de la base de datos. Dado que el test original permitía la respuesta *no sé*, la primera actividad fue depurar la base de datos de esta opción que, por su contenido actitudinal indeterminado, se hacía difícil de interpretar. Por ello, se eliminó a todas las personas que hubieran respondido *no sé* a 8 ó más ítems. Eso significó excluir a 88 casos, con lo que la muestra quedó compuesta finalmente por 1.023 personas.

Estudios sobre dimensionalidad. Uno de los supuestos de los modelos de la TRI utilizados es la unidimensionalidad del rasgo medido. Tomando en

cuenta que la definición nominal del constructo era compleja, se sospechó que el instrumento no cumpliría este requisito. Efectivamente, un análisis de componentes principales mostró que, si bien existía un componente principal dominante (17% de la varianza en los ítems), también aparecían otros dos factores marginales (que explicaban el 4% y 3,5% de dicha varianza, respectivamente).

Para eliminar esos factores marginales, se excluyó del instrumento a los ítems que, utilizando una extracción por componentes principales y otra por máxima verosimilitud, saturaban menos que 0,4 en el primer factor y correlacionaban fuertemente *(r* ≥ 0,3) con el segundo y tercer factor. Posteriormente se aplicó un modelo de ecuaciones estructurales con un solo factor común, eliminando aquellos ítems que presentaban regresiones cuadráticas estandarizadas menores que 0,20 sobre el factor común.

Esto significó eliminar 35 ítems, quedando 28 en la escala final, la que mostró un muy buen ajuste a un modelo factorial confirmatorio unifactorial (NFI[1] = 0,991; RMSEA = 0,038). Esta reducción del tamaño de la escala no eliminó ningún contenido del instrumento original sino que solo disminuyó el número de ítems con que se medía cada uno de ellos.

Estimación de los puntajes directos y utilización de la TCT. En esta fase se calcularon: (a) las puntuaciones directas de las personas, significando un mayor valor una mayor intolerancia (asignando 6 puntos a la respuesta que indicaba más intolerancia y 1 punto a la menos intolerante); (b) la posición de cada ítem utilizando el promedio de las respuestas; (c) los índices de discriminación de las afirmaciones; y (d) la consistencia interna de la escala por medio de alfa de Cronbach y Spearman-Brown.

Aplicación de los modelos de TRI dicotómicos. Para poder aplicar los modelos dicotómicos de la TRI, las respuestas de los individuos fueron previamente recodificadas de acuerdo al nivel de intolerancia que expresaban (las opciones 1, 2 y 3 se recodificaron como 0 o *baja intolerancia* y las alternativas 4, 5 y 6 como 1 o *alta intolerancia*). Se mantuvo como respuesta perdida la categoría no sé. La calibración de los parámetros de los modelos dicotómicos se realizó con el programa BILOG 3 (Mislevy & Bock, 1990), estimándose los parámetros de los ítems mediante el método de verosimilitud máxima (ML) y el nivel de rasgo de los individuos mediante el método modal bayesiano (MAP).

Aplicación de los modelos de TRI politómicos. La calibración de los ítems mediante los modelos MRG, MCP, MCPG y MRN se realizó utilizando el programa MULTILOG 6.0 (Thissen, 1991), obteniéndose una calibración de los parámetros de los ítems mediante el método ML y estimándose el rasgo de los individuos mediante el método MAP. Se mantuvo como respuesta perdida la categoría *no sé*. La calibración de los parámetros del MRGM se efectuó utilizando el programa PARSCALE (Muraki & Bock, 1997). Con él se realizó una estimación ML de los parámetros de los ítems y esperada a posteriori (EAP) del rasgo de los individuos. Finalmente, se utilizó el programa MODFIT 1.1 (Levine & Drasgow, 2001) para estimar el nivel de ajuste de los modelos politómicos.

Resultados

Calibraciones

Aplicación de la TCT. La posición de los ítems, calculada a partir del promedio de las respuestas (Tabla 1), nos lleva a sostener que hay bastante diversidad entre ellos, existiendo algunos que tienden a ser respondidos en alternativas de respuesta cercanas al polo de alta intolerancia (como el ítem 5) y otros en alternativas cercanas al polo de baja intolerancia (por ejemplo, el ítem 20). En general, predominan los ítems respondidos en alternativas de respuesta cercanas al polo de baja intolerancia (medias menores a 3).

La confiabilidad del test se calculó por consistencia interna, obteniéndose un valor alfa de Cronbach de 0,90 y una partición en mitades corregida por Spearman-Brown de 0,88. Por tanto, al menos desde la TCT, la escala tiene una alta consistencia interna.

Finalmente, la discriminación de los ítems se calculó por medio de la correlación r_{bp} de Pearson corregida entre cada ítem y la escala total, eliminando del total el ítem correspondiente. Los resultados presentados en la Tabla 1 muestran que todos los ítems presentan discriminaciones aceptables.

[1] NFI corresponde al índice de ajuste normado de Bentler y Bonnet y RMSEA corresponde al índice de error de aproximación cuadrático medio.

Tabla 1
Posición y Coeficientes de Discriminación de los Ítems

Ítem	Promedio de las respuestas	Correlación biserial puntual (r_{bp})	Ítem	Promedio de las respuestas	Correlación biserial puntual (r_{bp})
1	3,44	0,49	15	2,64	0,39
2	2,84	0,48	16	2,31	0,49
3	3,20	0,45	17	3,83	0,50
4	3,11	0,56	18	2,89	0,43
5	4,92	0,45	19	2,77	0,49
6	3,26	0,46	20	1,98	0,44
7	2,40	0,53	21	2,21	0,48
8	3,13	0,50	22	2,34	0,43
9	3,31	0,55	23	2,45	0,48
10	2,80	0,42	24	2,41	0,35
11	2,84	0,53	25	3,00	0,46
12	3,38	0,51	26	2,97	0,50
13	3,14	0,51	27	3,14	0,42
14	1,82	0,42	28	3,22	0,39

Aplicación de los modelos de TRI dicotómicos. En primer lugar, se calibraron los ítems con el modelo 1P. Este modelo tuvo un ajuste insatisfactorio, ya que 10 de las 28 afirmaciones presentaron χ^2 (9, N = 1023) mayores que 20, $p < 0,05$. Lo mismo ocurrió con la escala global: χ^2 (245, N = 1023) = 423,2, $p < 0,01$.

A continuación se aplicó el modelo 2P, el que presentó una concordancia con los datos significativamente mejor que el modelo logístico de un parámetro, ya que 26 de los ítems tuvieron un nivel satisfactorio de ajuste: χ^2 (9, N = 1023) menores que 16, $p > 0,05$. Lo mismo ocurrió con la escala global: χ^2 (241, N = 1023) = 281,7, p = 0,04. En otras palabras, sin mostrar un nivel perfecto, el modelo 2P ajusta suficientemente a los datos, indicándonos que es necesario considerar la existencia de diferentes niveles de discriminación de los ítems.

Con relación a los parámetros estimados por el modelo 2P, constatamos que, en promedio, el test tendría una discriminación baja (parámetro a = 0,70) y una posición algo superior al punto medio del continuo actitudinal (parámetro b = 0,65).

La baja discriminación puede deberse a que los ítems eran originalmente politómicos y fueron recodificados para calibrarlos con un modelo dicotómico. La dificultad algo superior indica que, según este modelo, el test medirá con menor error en los niveles medios y medio-superiores de la actitud que en sectores de baja intolerancia.

Aplicación de los modelos de TRI politómicos. Luego se calibraron las afirmaciones con el MRG. En el análisis de ítems individuales todos mostraron un grado de ajuste satisfactorio: los χ^2 (5, N = 1023) fueron menores que 2,2, $p > 0,05$. Además, prácticamente todas las díadas y tríadas de ítems se encuentran en niveles aceptables (tasas χ^2 / $g.l.$ < 4).

Dado este adecuado ajuste, estudiamos los parámetros estimados por el MRG en detalle. Como se puede observar en la Tabla 2, los parámetros de

Tabla 2
Parámetros Estimados MRG

	Parámetros					
	Discriminación	Posición				
	a	b_1	b_2	b_3	b_4	b_5
Promedio	1,10	-0,72	0,07	0,60	1,28	1,99
Desviación Estándar	0,17	0,71	0,74	0,73	0,75	0,68

discriminación de los ítems fueron en promedio bastante aceptables, pero los parámetros de posicionamiento se encuentran tan agrupados hacia el centro del rasgo que las opciones con más probabilidad de respuesta fueron las dos extremas en casi todo el continuo actitudinal.

En términos generales, como se muestra en la Figura 1, el instrumento parece ser más informativo en los sectores medios y medio-altos de la actitud que en los sectores bajos. Es así como la *función de información* estimada por el MRG, es decir, el grado en que el test permite medir en forma precisa a los individuos según su nivel de actitud, alcanza valores más altos en esos grados de intolerancia.

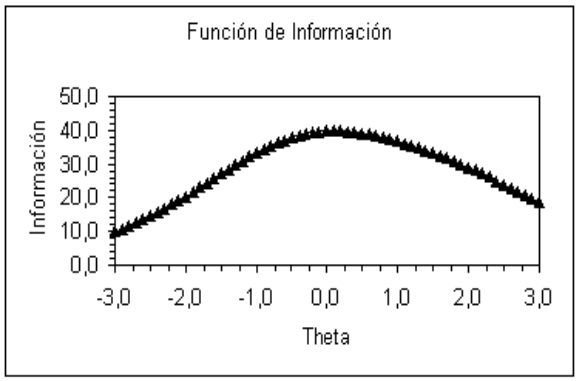

Figura 1. Función de información en el MRG.

A continuación se aplicó el MRGM. Este modelo no concordó con los datos, ya que ningún ítem presentó un ajuste satisfactorio: todos los χ^2 (5, N = 1023) fueron mayores que 55, $p < 0{,}01$.

Por su parte, tampoco el MCP presentó un buen ajuste. Es así como solo un ítem tuvo un χ^2 adecuado (χ^2 [5, N = 1023] = 12,68, $p > 0{,}05$), mientras que el resto estuvo muy lejos de esa situación (27 χ^2 [5, N = 1023] fueron mayores que 23, $p < 0{,}01$).

Luego se aplicó el MCPG, el cual tampoco fue satisfactorio. Solo un ítem tuvo una concordancia adecuada (27 χ^2 [5, N = 1023] fueron mayores que 19, $p < 0{,}01$).

Para finalizar, se calibraron los ítems con el MRN, presentando una buena correspondencia con los datos. En el análisis de ítems individuales todos mostraron un nivel de ajuste satisfactorio: los χ^2 (5, N = 1023) fueron menores que 2,0, $p > 0{,}05$. Además, prácticamente todas las díadas y tríadas de afirmaciones se encuentran en niveles aceptables (tasas χ^2 / *g.l.* < 4).

No obstante, se obtuvieron parámetros a_1 y a_6 negativos, lo que implicaría que las personas con mayor nivel de intolerancia tienden a no escoger las respuestas que indican más intolerancia, sino las que contienen menos intolerancia y viceversa.

Lo anterior representa un desajuste en el orden de las opciones de respuesta. No encontramos evidencias que permitan sostener un cambio de orden en las alternativas de respuesta, por lo que podemos hipotetizar, a la luz de investigaciones anteriores (De Ayala & Sava-Bolesta, 1999), que los parámetros estimados por el MRN en este caso concreto no permiten interpretar correctamente las respuestas, ya que no disponemos del tamaño de muestra necesario para estimar sin un excesivo error los parámetros de las alternativas extremas.

Por todo ello, a continuación solo se hará referencia a los otros dos modelos que se corresponden con los datos: 2P y MRG. Se trabajó prioritariamente con este último, que aúna un buen ajuste, supuestos razonables, aprovechamiento de la información politómica y resultados interpretables.

Escalamiento

Uno de nuestros principales objetivos era determinar si existía correspondencia entre el escalamiento generado por los modelos de la TRI y el obtenido mediante las puntuaciones directas. La correlación lineal de Pearson entre las puntuaciones directas y el modelo 2P fue 0,95 (656, $p < 0{,}01$), y con el modelo MRG fue 0,98 (656, $p < 0{,}01$).

En la Figura 2 confirmamos el carácter lineal de la asociación para el caso del MRG.

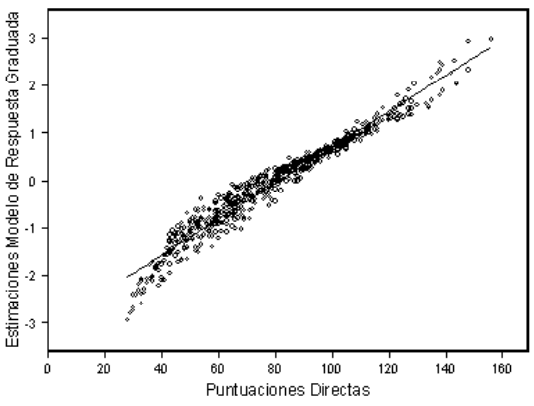

Figura 2. Relación entre puntuaciones directas y estimación de actitud en el MRG.

También fue lineal la relación entre el escalamiento producido por el modelo 2P y el MRG: la correlación de Pearson entre ellos fue 0,96 (1023, $p < 0{,}01$).

Parámetros

Discriminación de los ítems. También hay una alta relación lineal entre la discriminación de los ítems estimada utilizando la TCT y aplicando los modelos de la TRI. La correlación lineal de Pearson entre las puntuaciones directas y el modelo 2P fue 0,71 (28, $p < 0{,}01$), y con el modelo MRG, 0,75 (28, $p < 0{,}01$).

Coincidente con lo anterior, en la Figura 3 se muestra que la relación entre las pendientes definidas por el MRG y los niveles de discriminación de la TCT es aproximadamente lineal, aunque con una más amplia dispersión respecto a la recta de regresión.

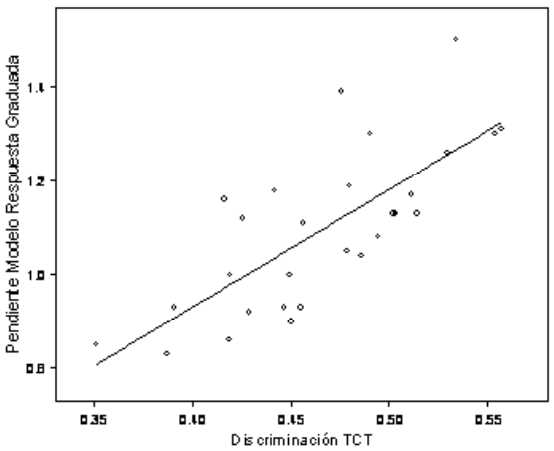

Figura 3. Relación entre la discriminación en la TCT y la pendiente del MRG.

Posicionamiento de los ítems. Al calcular puntuaciones directas, se utilizó el promedio de las respuestas de los individuos a cada ítem como indicador para determinar el posicionamiento de cada afirmación. Comparar este indicador con el estimado por el modelo 2P es simple, pues este presenta tal parámetro, pero no así con el MRG, que dispone de cinco parámetros de posición como se ha mostrado en la Tabla 2. Por ello, en este caso se utilizaron como indicadores aproximados de la posición del ítem el parámetro b_3 (que indica el punto del continuo actitudinal en que es tan probable responder a las tres primeras alternativas como a las tres segundas) y el promedio de los cinco parámetros *b*.

La relación lineal entre estos parámetros fue casi perfecta. La correlación lineal de Pearson entre el posicionamiento de los ítems en la TCT y el indicador de posición del modelo 2p fue -0,97 (28, $p < 0{,}01$), con el parámetro b_3 fue -0,97 (28, $p < 0{,}01$) y con la media de parámetros *b* fue -0,98 (28, $p < 0{,}01$).

A diferencia del posicionamiento estimado por las puntuaciones directas y el modelo 2P, el MRG nos permite conocer otra característica de los ítems: el grado en que sus alternativas evalúan un sector amplio o estrecho del continuo actitudinal.

Si un ítem tiene sus parámetros *b* agrupados en un pequeño sector del continuo, la precisión con que medirá ese tramo será muy alta, pero a costa de aportar poca información en el resto de los niveles de la actitud. Por el contrario, un ítem con sus parámetros *b* extendidos a lo largo del continuo contribuirá a evaluar más homogéneamente el rasgo.

En términos generales, los parámetros *b* del MRG se encuentran agrupados en el centro del continuo actitudinal, poniendo en duda la pertinencia de haber utilizado cuatro alternativas de respuesta intermedias que aportan poco a la información total obtenida por cada ítem.

Por otro lado, analizando detalladamente los parámetros *b* de cada ítem, podemos observar algunas diferencias en su amplitud, por lo cual se los clasificó en tres grupos: el cuartil de afirmaciones de parámetros *b* de rango más estrecho, los dos cuartiles intermedios y el cuartil con parámetros *b* más amplios.

Estudiando el contenido de dichos ítems, parece haber una tendencia a que los parámetros amplios se produzcan principalmente frente al tema de las diferencias sociales y de las transgresiones culturales en materia de sexualidad, vestimenta o consumo de drogas. Por su parte, los parámetros *b* estrechos parecen generarse en el campo de las opiniones políticas y la discriminación hacia los jóvenes, homosexuales o mujeres.

En síntesis, si bien el posicionamiento de los ítems estimado por las puntuaciones directas no es muy distinto de los parámetros de localización obtenidos en los modelos 2P y MRG, este último nos permite detectar aquellas afirmaciones que miden segmentos específicos del continuo actitudinal, por lo que un análisis de su contenido puede permitir redactar nuevos ítems que mejorarían la calidad del

instrumento en niveles de actitud donde la estimación de habilidad se realiza con mayor error.

Consistencia Interna y Error de Estimación

El procedimiento derivado de la TCT solo permite estimar el nivel global de consistencia interna del test, el cual fue bastante aceptable en el presente estudio (alfa de Cronbach = 0,90). En cambio, el MRG nos permite estimar tanto el grado de información que se obtiene por nivel de actitud como el error con que se ha medido a cada individuo.

En la Figura 4 se muestra que el error de estimación de la actitud de los participantes es mayor a niveles bajos de actitud que a niveles medios y altos.

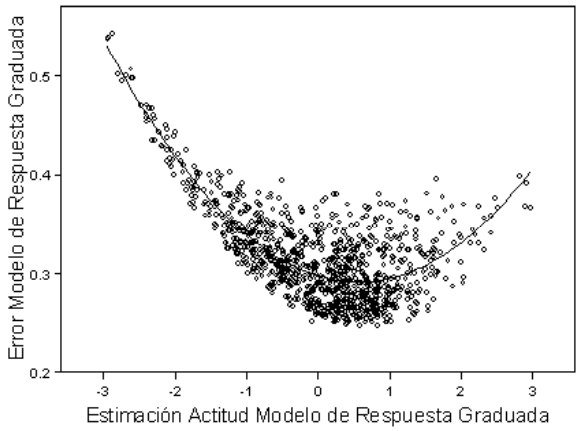

Figura 4. Relación entre la estimación de habilidad de las personas y el error de estimación en el MRG.

Podemos sostener que el instrumento resulta bastante confiable al evaluar la actitud de las personas con niveles medios y altos de intolerancia, pero mucho menos al estimar el rasgo de los individuos con menor nivel. Esto supone que si el objetivo del test fuera evaluar a dichas personas, se debería incluir mayor cantidad de ítems que discriminen en este nivel actitudinal.

Otra importante información que podemos obtener de un modelo como el MRG es el grado en que aparecen patrones de respuesta *atípicos* o *aberrantes*, es decir, contradictorios. Como se puede observar en la Figura 4, si bien en nuestros datos hay bastante heterogeneidad en el nivel de error con que son medidas las personas a iguales niveles de actitud, no se observan personas con errores desusadamente altos para su nivel de rasgo, por lo que se puede descartar la existencia de patrones aberrantes de respuesta.

Asociación con Variables Adicionales

Originalmente los puntajes de las personas en la escala de intolerancia fueron estimados sumando las respuestas a cada uno de los 63 ítems iniciales. Paralelamente, nuestra investigación generó tres nuevos puntajes de la actitud de las personas: (a) los puntajes directos de los 28 ítems retenidos en el test unidimensional final, (b) la estimación del rasgo de los individuos en el modelo 2P y (c) la estimación de actitud de los participantes en el MRG.

En la Tabla 3 se puede observar la asociación entre estas cuatro formas de estimación de la actitud y las seis variables adicionales que consideraba la investigación. Se puede apreciar que utilizar cualquiera de las cuatro estimaciones de habilidad nos llevaría a prácticamente las mismas conclusiones respecto de la asociación de la intolerancia con las seis variables adicionales.

Esta alta correspondencia se explica por la fuerte correlación lineal entre las puntuaciones directas originales y nuestras tres estimaciones de habilidad. La correlación lineal de Pearson entre las puntuaciones directas originales y las estimaciones de habilidad en el modelo 2P fue 0,91 (656, $p < 0,01$), con las estimaciones del MRG fue 0,94 (656, $p < 0,01$) y con las puntuaciones directas unidimensionales fue 0,95 (656, $p < 0,01$).

El carácter lineal de la relación se puede observar en la Figura 5, en la que se compara la estimación de habilidad en el MRG y las puntuaciones directas originales.

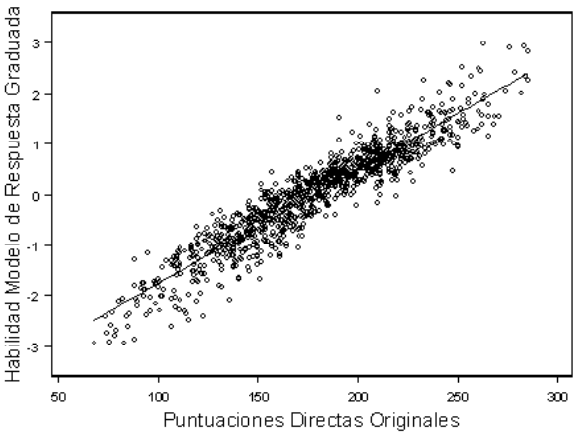

Figura 5. Relación entre puntuaciones directas originales y estimación de actitud en el MRG.

Tabla 3
Asociación entre Variables de Identificación Sociodemográfica y Estimaciones de la Actitud de los Individuos

Variables Adicionales	Puntuaciones Directas Originales	Puntuaciones Directas Unidimensionales	Estimación Actitud Modelo 2 Parámetros	Estimación Actitud Modelo Respuesta Graduada
Sexo	t = -0,23 p = 0,82	t = 0,95 p = 0,43	t = -0,24 p = 0,87	t = -0,57 p = 0,60
Religión	f = 4,24 p = 0,01	f = 4,34 p = 0,01	f = 3,41 p = 0,03	f = 3,81 p = 0,02
Edad	r(p) = 0,28 p < 0,01	r(p) = 0,26 p < 0,01	r(p) = 0,24 p < 0,01	r(p) = 0,26 p < 0,01
Educación	r(s) = -0,46 p < 0,01	r(s) = -0,49 p < 0,01	r(s) = -0,50 p < 0,01	r(s) = -0,47 p < 0,01
Nivel Socioeconómico	r(s) = 0,30 p < 0,01	r(s) = 0,36 p < 0,01	r(s) = 0,36 p < 0,01	r(s) = 0,33 p < 0,01
Posición Política	r(s) = 0,37 p < 0,01	r(s) = 0,30 p < 0,01	r(s) = 0,26 p < 0,01	r(s) = 0,29 p < 0,01

t: t de Student
f: f de Fisher
r(p): Correlación de Pearson
r(s): Correlación de Spearman

Discusión

Calibrados siete modelos dicotómicos y politómicos de TRI, solo presentaron un ajuste satisfactorio tres de ellos: el modelo de 2P, el MRN y el MRG. Esta dificultad en el ajuste de modelos de la TRI a test no cognitivos es coincidente con lo encontrado en otras investigaciones (Chernyshenko et al., 2001). Una posible explicación de esta dificultad es que este tipo de investigaciones se ha desarrollado hasta ahora principalmente sobre test diseñados con la TCT y no con la TRI, lo que hace más difícil su ajuste.

Por otro lado, se comprueba también lo descrito en anteriores estudios (Barbero, Prieto, Suárez & San Luis, 2001) respecto de la similitud de los parámetros de los individuos y de los ítems calibrados por modelos de la TRI y la TCT. Un corolario de esta semejanza es que la asociación entre los puntajes de los individuos y algunas variables sociodemográficas fue similar para todas las estimaciones actitudinales.

Para encontrar las ventajas de aplicar modelos de la TRI tenemos que centrarnos en el tema de la información respecto del instrumento que nos permiten obtener estos modelos. En nuestro estudio el MRG nos permitió:

1. Conocer el error de estimación de la actitud para cada nivel del rasgo y para cada persona.
2. Realizar un análisis detallado del contenido de los ítems en función de sus parámetros de localización y discriminación, lo que permitiría mejorar la calidad del instrumento, agregando afirmaciones pensadas en ser más informativas en segmentos específicos del continuo actitudinal.
3. Conocer la información que aporta cada alternativa de respuesta, lo que nos permite sugerir la reducción del número de alternativas en futuras aplicaciones del test. Nuestros datos respaldan estudios previos que afirman que las opciones intermedias son poco informativas (González-Romá & Espejo, 2003; Hernández, Espejo, González-Romá & Gómez, 2001) y que cuatro alternativas de respuesta son suficientes para modelos de TRI politómicos (Gray-Litter et al., 1997; Hernández et al., 2000).

Que algunos modelos no hayan ajustado a los datos mientras que otros muy similares sí lo hayan hecho nos proporciona información interesante respecto del instrumento. Así, el desajuste del modelo 1P nos indica que no podemos considerar iguales los parámetros de discriminación de los ítems. Por su parte, la poca concordancia con los datos del modelo MRGM nos permite inferir que la distancia entre las alternativas de respuesta no es constante a lo largo de todos los ítems.

Uno de los resultados inesperados del estudio fueron los parámetros estimados por el MRN. En

la literatura existen pocos ejemplos de la aplicación de este modelo a escalas tipo Likert. En algunas de estas investigaciones el orden de las alternativas de respuesta se ha mantenido dentro de lo esperado (González-Romá & Espejo, 2003; Hernández et al., 2001;) y en otras se ha detectado desorden en las opciones (Espejo & González, 1999). Esta situación puede ser explicada tanto por la presencia de un desorden real de las opciones de respuestas como por errores de estimación de los parámetros por parte del MRN. En este caso podemos sostener que para recuperar correctamente los parámetros reales aplicando el MRN, en condiciones de distribución normal del rasgo, se debe disponer de una tasa 20:1 de individuos por cada parámetro a estimar (De Ayala & Sava-Bolesta, 1999), por lo que hipotetizamos que en este caso los parámetros que hemos obtenido con el MRN no son adecuados.

Reconociendo que las conclusiones de esta investigación son producto de un estudio de un solo caso y, por tanto, de limitada generalización, ¿qué podemos decir respecto de nuestra pregunta sobre el aporte que pueden hacer los modelos de la TRI en una situación como la estudiada?

En síntesis, la utilización de la TRI resultó particularmente útil tanto en la determinación del error con que medimos a nuestros entrevistados como en la descripción de las características psicométricas del instrumento que hemos aplicado, permitiéndonos formular sugerencias para su mejoramiento.

Como contrapartida, estos modelos no modificaron significativamente el escalamiento de los individuos ni cambiaron el grado de asociación entre la actitud y las variables adicionales. ¿Se justifica, entonces, la utilización de modelos de TRI politómicos en una investigación actitudinal, a pesar que su aplicación exige mayores recursos? Si bien la respuesta a esta pregunta dependerá de cada investigador, este estudio ha mostrado que estos modelos tienen un aporte relevante que hacer al mejoramiento de los instrumentos usados para la medición de actitudes.

Además, no podemos olvidar que la invarianza de las estimaciones que hacemos por medio de estos modelos politómicos y el hecho que estas se encuentren en una métrica común, está permitiendo el diseño de TAIs más eficientes y avanzar en los procesos de equiparación y detección del funcionamiento diferencial en escalas actitudinales o de personalidad.

Finalmente, el desarrollo de nuevos modelos politómicos no paramétricos o multidimensionales facilitará el ajuste de los datos a los modelos utilizados y la estimación de sus parámetros con muestras más pequeñas, potenciando el uso futuro de modelos politómicos de la TRI en la medición aplicada de constructos actitudinales y de personalidad.

Referencias

Aymerich, J. (2001). *Dimensiones temáticas de intolerancias y discriminaciones* (Documento de Trabajo del Departamento de Sociología Nº 3). Santiago, Chile: Universidad de Chile, Departamento de Sociología.

Barbero, M. I., Prieto, P., Suárez, J. C. & San Luis, C. (2001). Relaciones empíricas entre los estadísticos de la Teoría Clásica de los Tests y los de la Teoría de Respuesta a los Ítems. *Psicothema, 13,* 324-329.

Birnbaum, A. (1957). *Efficient design and use of tests of ability for various decision-making problems* (Series Report Nº 58-16, Project Nº 7755-23). Universal City, TX: US Air Force, Randolph Air Force Base, School of Aviation Medicine.

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

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F. & Williams, B. (2001). Fitting Item Response Theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*, 523-562.

De Ayala, R. J. & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement, 23,* 3-19.

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Espejo, B. & González, V. (1999, Octubre). *El orden de las alternativas de respuesta en escalas tipo Likert: un estudio mediante modelos de la Teoría de Respuesta al Ítem*. Ponencia presentada en el VI Congreso de Metodología de las Ciencias Sociales y de la Salud, Oviedo, España.

Gómez, J., Hidalgo, M. D. & Tomás-Sábado, J. (2007). Using polytomous item response models to assess death anxiety. *Nursing Research, 56,* 89-96.

González-Romá, V. & Espejo, B. (2003). Testing the middle response categories «not sure», «in between» and «?» in polytomous items. *Psicothema, 15,* 278-284.

Gray-Litter, B., Williams, V. S. & Hancock, T. D. (1997). An Item Response Theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 23,* 443-451.

Hernández, A., Espejo, B., González-Romá, V. & Gómez, J. (2001). Escalas de respuesta tipo Likert: ¿es relevante la alternativa "indiferente"? *Metodología de Encuestas, 3,* 135-150.

Hernández, A., Muñiz, J. & García, E. (2000). Comportamiento del modelo de respuesta graduada en función del número de categorías de la escala. *Psicothema, 12* (Supl. 2), 288-291.

Hol, A. M., Vorst, H. C. & Mellenbergh, G. J. (2005). A randomized experiment to compare conventional, computerized, and computerized adaptive administration of ordinal polytomous attitude items. *Applied Psychological Measurement, 29,* 159-183.

Javaras, K. N. & Ripley, B. D. (2007). An "unfolding" latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association, 102,* 454-463.

Kim, S., Cohen, A. S., Alagoz, C. & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items.

Journal of Educational Measurement, 44, 93-116.

Lai, J., Cella, D., Chang, C., Bode, R. K. & Heinemann, A. W. (2003). Item banking to improve, shorten and computerize self-reported fatigue: An illustration of steps to create a core item bank from the FACIT-Fatigue Scale. *Quality of Life Research, 12,* 485-501.

Lee, G., Kolen, M. J., Frisbie, D. A. & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement, 25,* 357-372.

Levine, M. & Drasgow, F. (2001). *MODFIT 1.1.* Urbana, IL: University of Illinois at Urbana-Champaign.

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

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

McKinley, R. (1989). An introduction to Item Response Theory. *Measurement and Evaluation in Counseling and Development, 22,* 37-57.

Mislevy, R. J. & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software International.

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

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 69-71.

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

Muraki, E. & Bock, R. D. (1997). *PARSCALE, IRT item analysis and test scoring for rating-scale data*. Chicago: Scientific Software International.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Roberts, J. S. & Laughlin, J. E. (1996) A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20,* 231-255.

Rojas A. J. & Lozano, O. M. (2005). Application of an IRT polytomous model for measuring health related quality of life. *Social Indicators Research, 74,* 369-394.

Samejima, F. (1969). Estimation of latent ability using a response pattern of grades scores. *Psychometrika Monograph Nº 17*.

Santisteban, C. & Alvarado, J. M. (2001). *Modelos psicométricos*. Madrid: Universidad Nacional de Educación a Distancia.

Stout, W. (2001). Nonparametric Item Response Theory: A maturing and applicable measurement modeling approach. *Applied Psychological Measurement, 25,* 300-306.

Thissen, D. (1991). MULTILOG: *Multiple, categorical item analysis and test scoring using Item Response Theory*. Chicago: Scientific Software International.

Van Rijn, P. W., Eggen, T. J., Hemker, B. T. & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 26,* 393-411.

Fecha de recepción: Marzo de 2008.
Fecha de aceptación: Octubre de 2008.