ANUARIO DE PSICOLOGÍA
Núm. 43 - 1989 (4)

ROBUSTEZ DE LAS ESTIMACIONES DEL MODELO DE RASCH EN PRESENCIA DE ACIERTOS AL AZAR Y DISCRIMINACIÓN VARIABLE DE LOS *ITEMS*

JOSÉ MUÑIZ
Universidad de Oviedo
JANE ROGERS
HARIHARAN SWAMINATHAN
Universidad de Massachusetts

José Muñiz
Departamento de Psicología
Universidad de Oviedo
P. de Asturias

Los Modelos de Teoría de Respuesta a los *Items* (TRI), otrora denominados Modelos o Teoría del Rasgo Latente, aportan soluciones novedosas para algunos de los problemas tecnológicos que tenía planteados la Teoría Clásica de los Tests (TCT) y que parecían intratables dentro del marco de referencia clásico. Ello ha obligado a reanalizar y reconstruir bajo la nueva óptica áreas enteras de la Psicometría, tales como la Evaluación del Sesgo de los Tests, la Equiparación de las puntuaciones (*Equating*), los Bancos de *Items*, Tests Adaptados al sujeto, Tests Computerizados, Tests Referidos al Criterio, etc., convirtiéndose la TRI en la línea de investigación psicométrica predominante, como fácilmente se puede comprobar echando una ojeada a los contenidos de las revistas y otras publicaciones del área.

Exposiciones sistemáticas de la TRI pueden consultarse en Andrich (1988), Hambleton (1983), Hambleton y Swaminathan (1985), Hulin, Drasgow y Parsons (1983), Lord (1980), Lord y Novick (1968), Rasch (1960) o Wright y Stone (1979), aquí sólo se mencionarán aquellos aspectos de los *Modelos Logísticos* de uno (Modelo de Rasch), dos y tres parámetros (Birnbaum, 1968; Lord, 1980) que se juzguen relevantes para clarificar el problema abordado. Las dos grandes *aportaciones* de los modelos de TRI pueden cifrarse en la *invarianza de los parámetros* de los *items* respecto de los sujetos y la *invarianza de las mediciones* respecto del instrumento utilizado. En la Teoría Clásica las propiedades de los *items* dependen del tipo de sujetos utilizados para establecerlas, así, por ejemplo, un mismo *item* tendrá distinto Índice de Dificultad en función de que la muestra utilizada para calcularlo esté compuesta de sujetos competentes o incompetentes, en el primer caso el *item* resultará fácil y en el segundo difícil. Por contra, en la TRI los parámetros que definen los *items* no dependen de los objetos medidos, los sujetos, lo cual parece imprescindible para la obtención de mediciones objetivas (Wright y Stone, 1979). En segundo lugar, cuando en la TCT se mide una variable con distintos tests se obtienen resultados que no están en la misma escala, no son comparables sin más, es decir, las mediciones no son independientes del instrumento de medida utilizado. Por ejemplo, si se mide la Comprensión Verbal con dos tests de Sinónimos distintos, se obtendrán resultados diferentes, no directamente comparables, se desconoce la función de equivalencia entre uno y otro. Ello imposibilita, entre otras cosas, la comparación cabal entre sujetos que hayan respondido a tests distintos midiendo la misma variable. La TRI también resuelve este problema, las mediciones ya no dependerán del test utilizado, los resultados se expresan en una métrica común que no es la de ningún test en

* Esta investigación fue realizada durante la estancia en la Universidad de Massachusetts con una beca del Comité Conjunto Hispano Norteamericano para la Cooperación Cultural y Educativa, quede constancia de mi agradecimiento.

particular, tests distintos generan la misma estimación de la competencia de los sujetos. Hay que señalar, no obstante, que en el marco de la TCT se habían ideado algunos remedios prácticos para mitigar las dos limitaciones citadas, pero la solución teórica general llegará con la TRI. *Otra ventaja* importante de la TRI es que el Error Típico de Medida no es el mismo para todos los sujetos, depende de su nivel en la variable medida, el test será más preciso para unos niveles que para otros, lo que se operativiza mediante la Función de Información.

Las implicaciones son notables, pues ello permitirá seleccionar el test que sea más eficaz para medir al nivel que estemos interesados, ajustándose al sujeto para minimizar los errores de medida, algo parecido al establecimiento de los umbrales en algunos métodos psicofísicos. El precio a pagar por estas ventajas de los modelos de TRI es que se basan en unos supuestos bastante más restrictivos que los exigidos por la TCT. Los modelos mejor estudiados hasta la fecha, los Logísticos de uno, dos y tres parámetros, asumen la *unidimensionalidad de los items*, esto es, que todos miden la misma variable, o, en otras palabras, que las respuestas de los sujetos a los *items sólo* dependen de su nivel en la variable medida. La unidimensionalidad implica Independencia Local, a saber, que la respuesta a un *item* no influye en la respuesta a otro, nótese que de lo contrario se inflingiría el que la conducta ante los *items* sólo depende de la variable medida, según lo predicado por la unidimensionalidad. Se han propuesto diversos modelos multidimensionales, probablemente más acordes con las situaciones reales (Bock y Aitkin, 1981; Samejima, 1974; Thissen Steinberg, 1984; Whitely, 1980), pero lo que ganan en realismo lo pierden en complejidad, y, sobre todo, en precisión a la hora de estimar los parámetros, con problemas de estimación todavía no bien resueltos. Una clasificación de los modelos de TRI atendiendo a varios criterios, incluida la dimensionalidad, puede consultarse en Thissen y Steinberg (1986). Los trabajos encaminados a evaluar la robustez de los modelos a violaciones del supuesto de unidimensionalidad son abundantes (Ansley y Forsyth, 1985; Drasgow y Parsons, 1983; Greaud, 1988; Harrison, 1986; McKinley y Mills, 1985; Reckase, 1979; Wollenberg, 1982; Yen, 1984) y las conclusiones varias, siendo éste uno de los puntos calientes de la investigación actual. Como es fácil de imaginarse, el asunto es cuestión de grado, a medida que la varianza explicada por el primer factor aumenta, más perfecta es la unidimensionalidad, luego la gravedad de las violaciones disminuye; pero el acuerdo entre los investigadores acerca de cuánta varianza debe explicar el primer factor para poder aplicar los modelos con ciertas garantías, o, qué tipos de índices son más adecuados para su evaluación, dista mucho de ser unánime. Hattie (1985), por ejemplo, cita 87 índices utilizados en la literatura para evaluar la unidimensionalidad. El criterio general más utilizado sigue siendo que la proporción entre la varianza explicada por el primer factor y la explicada por el segundo sea «razonablemente» alta (Lumsden, 1961, 1976; Lord, 1980), si bien otros criterios con mayor rigor estadístico son deseables, aunque la significación estadística no siempre vaya acompañada de la psicológica.

El punto clave en el desarrollo de los modelos se centra en la estimación correcta y precisa de los parámetros. Utilizando datos simulados se ha investigado ampliamente la recuperación de los parámetros bajo diferentes condiciones,

manipulando aspectos tales como el número de sujetos, su distribución en la variable medida, distribuciones de los parámetros, número y tipos de *items*, etc., (Divgi, 1986; Hambleton y Cook, 1983; Hulin, Lissak y Drasgow, 1982; Kolen, 1981; Lord, 1975, 1983, 1986; McKinley y Mills, 1985; Mislevy, 1986; Ree, 1979; Swaminathan y Gifford, 1982, 1983, 1985, 1986; Thissen y Wainer, 1982; Wingersky y Lord, 1984; Yen, 1987). Véase Baker (1987) para una buena revisión y estado actual de la cuestión. Los tres procedimientos de estimación más utilizados han sido el de *Máxima Verosimilitud Conjunta*, implementado en los programas de computador LOGIST (Wingersky, Barton y Lord, 1982) y BICAL (Wright, Mead y Bell, 1979), *Máxima Verosimilitud Marginal*, implementado en BILOG (Mislevy y Bock, 1984) y *Estimaciones Bayesianas* (Swaminathan y Gifford, 1982, 1983, 1985, 1986). Un análisis comparativo de los tres procedimientos puede verse en Lord (1986). La estimación del parámetro b (Dificultad) está bien resuelta teóricamente y se lleva a cabo con precisión, pero no ocurre lo mismo con la de a (Discriminación), y, sobre todo, con la de c (Aciertos al Azar), especialmente cuando las muestras no son muy numerosas (menos de 50 *items* y 500-1000 sujetos aproximadamente). Kolen (1981), Ree (1979) y Thissen y Wainer (1982), entre otros, presentan datos bastante convincentes acerca de la imprecisión de las estimaciones de c, siendo ya clásicas por otra parte las objeciones de Wright (Wright, 1977 a y b; Wright y Stone, 1979) al uso de los modelos logísticos de dos y tres parámetros. Esta relativa imprecisión en la estimación de a y c con muestras pequeñas plantea la interesante cuestión de si en estos casos sería aconsejable usar el modelo de Rasch que sólo conlleva la estimación del parámetro b para cada *item*. Lord (1983) responde que sí, que muestras pequeñas (10-15 *items* y menos de 100-200 sujetos) justifican el uso del modelo de Rasch frente al de dos parámetros. Si se tiene en cuenta que en la práctica lo más corriente es disponer de pocos sujetos y no excesivos *items*, y si a ello se añade el atractivo teórico del Modelo de Rasch, su sencillez, que lo hace comprensible para no especialistas, y el fácil manejo del programa BICAL en el que está implementado, se tienen los ingredientes que explican la omnipresencia y uso masivo del modelo en la práctica. Para 1978 Rentz y Rentz (1978) dan ya cuenta de más de 400 referencias sobre el modelo sólo en Estados Unidos, hoy, una década después, a buen seguro que se cuentan por varios miles.

Como es bien sabido en el Modelo de Rasch la probabilidad de acertar un *item* viene dada por la función Logística:

$$p(\theta) = 1/[1+\exp\{-Da(\theta-b_i)\}]$$

donde:

- θ es la variable medida.
- b_i es la Dificultad del *Item* i.
- a es una constante que representa el poder discriminativo común para todos los *items*.
- D es una constante de escalamiento que para el valor 1.7 aproxima la Función Logística a la Normal.
- exp es la base de los logaritmos Neperianos (2.71).

La probabilidad de acertar un *item* sólo depende de su dificultad (b) y del nivel del sujeto en la variable medida (θ). Por el contrario, el Modelo Logístico de *dos* Parámetros añade para caracterizar los *items* otro parámetro, el Índice de Discriminación (a_i):

$$P(\theta) = 1/[1+\exp\{-Da_i(\theta-b_i)\}]$$

y el Modelo de *tres* parámetros incluye además los Aciertos al Azar (c_i):

$$P(\theta) = c_i + (1-c_i)/[1+\exp\{-Da_i(\theta-b_i)\}]$$

Este uso tan extendido del Modelo de Rasch trae como consecuencia que gran parte de los tests a los que se aplica, la mayoría, están compuestos por *items* de elección múltiple, en los que la probabilidad de acertar al azar es relativamente alta. Por ejemplo, si un *item* consta de dos alternativas (una correcta), la probabilidad de acertarlo al azar cuando no se sabe nada será 0.50, si tiene tres 0.33, si cuatro 0.25, etc. [La probabilidad de acertar un *item* al azar no es estrictamente la definición del parámetro c, pero puede considerarse como una buena estimación por exceso: $c = p(\theta = -\infty)$]. Por tanto, al aplicar el Modelo de Rasch a estos tests, lo cual se hace rutinariamente, se viola claramente la condición exigida por el modelo de que c sea cero, especialmente si las omisiones no se permiten o se desaconsejan, caso bien frecuente. Otra violación clara que se produce al aplicar el Modelo de Rasch a tests de Elección Múltiple (y otros) es la asunción de que todos los *items* tienen el mismo Índice de Discriminación. Como señala Traub (1983), asumir que no existen aciertos al azar y que el poder discriminativo de los *items* es constante va contra el sentido común y 80 años de evidencia empírica. La incidencia de c en el ajuste del modelo de Rasch ya fue subrayada por McKinley y Mills (1985), Reckase (1979), Wainer y Wright (1980) y Yen (1981), mientras que los valores de a parecen afectar menos la precisión de las estimaciones (Hambleton y Traub, 1971; Hambleton y Cook, 1983). Van de Vijver (1986) llevó a cabo un trabajo en el que manipuló, como se hará aquí, los valores de c y a, confirmando la escasa incidencia de a sobre la precisión de las estimaciones y la algo más relevante de c cuando la precisión se evalúa como las discrepancias simulado/estimado de la Curva Característica en vez de a nivel correlacional. El reanálisis de los datos de Dinero y Haertel (1977) llevado a cabo por Van de Vijver (1986) también apunta en la dirección de la escasa incidencia de a.

En suma, con *items* de Elección Múltiple parece más que probable que existan Aciertos al Azar ($c \neq 0$) y que sus Índices de Discriminación no sean iguales ($a \neq K$), por lo que sería teóricamente desaconsejable a priori el uso del Modelo de Rasch, habría que recurrir al más complejo de tres parámetros. Sin embargo, por las razones ya citadas, es habitual utilizar el modelo de Rasch en estas condiciones, haciéndose necesario conocer con exactitud cómo se ven afectadas las estimaciones de los parámetros en tales circunstancias. Ese será el *objetivo* central de la presente investigación, evaluar la robustez de las estimaciones de b y θ con el Modelo de Rasch cuando se dan Aciertos al Azar ($c \neq 0$) y los Índices de Discriminación de los *Items* no son iguales ($a \neq K$). Los resultados, aparte del interés

teórico y estadístico, podrían tener serias implicaciones para el uso tan extendido del Modelo de Rasch con tests de elección múltiple. Tal vez parezca chocante teóricamente este empleo de un modelo, el de Rasch, en unas condiciones en las que es altamente probable que no se cumplan sus asunciones, existiendo además otro, el de tres parámetros, que encajaría perfectamente. Pero, como ya se ha señalado, la estimación de a y c en el modelo de tres parámetros conlleva cierta imprecisión, especialmente si las muestras son pequeñas y los *items* no excesivos como ocurre a menudo en la práctica. Por tanto, no es descabellado pensar que debido al ruido introducido por la estimación imprecisa de a y c, quizás el uso (más apropiado) del modelo de tres parámetros no produzca una mejora significativa respecto del uso (menos apropiado) del modelo de Rasch, en cuyo caso el principio de parsimonia aliado al sentido común aconsejaría el modelo más simple. Este argumento, a menudo invocado por los partidarios del modelo de Rasch, aunque razonable, no está sobrado de evidencia empírica sustentadora, si bien la existente ya citada parece apoyarlo; trataremos por nuestra parte de añadir alguna en una u otra dirección. El análisis detallado del objetivo propuesto se llevará a cabo desglosado en tres estadios. En primer lugar se estudiarán sistemáticamente las relaciones entre los valores de c (Aciertos al Azar) y la precisión de las estimaciones del Modelo de Rasch. En segundo lugar se verá cómo afecta a las estimaciones la presencia de Índices de Discriminación (a) variable de los *items*. Finalmente se evaluará la influencia sobre las estimaciones de la interacción de ambos factores, a y c. Además, se comprobará en cada caso si el uso de los modelos logísticos de dos y tres parámetros en vez del modelo de Rasch introduce ganancias significativas en la precisión de las estimaciones de los parámetros y por ende en el ajuste del modelo. Todo ello se llevará a cabo con datos simulados.

Simulación de los datos

La lógica general de la investigación consistirá en generar datos con parámetros conocidos para diferentes valores de c y a y examinar la precisión de su recuperación por los modelos logísticos de 1, 2 y 3 parámetros. Dado que el interés se centra en evaluar la influencia de los Aciertos al Azar (c) y del poder discriminativo (a) en la precisión de las estimaciones del modelo de Rasch, se fijaron seis valores para c y dos para a del siguiente modo:

Valores de c:

* 0.00 Inexistencia de Aciertos al Azar para todos los *items*.
* 0.00-0.50 Los valores c de los *items* están comprendidos entre 0.00 y 0.50 con Distribución Uniforme. Ello representaría aproximadamente los valores típicamente encontrados en las aplicaciones empíricas de los modelos.
* 0.50 Todos los *items*. Raramente, si alguna vez, se encuentran valores tan elevados de c, pero sería pensable teóricamente para *items* con dos alternativas, por ejemplo.

* 0.33 Todos los *items.*
* 0.25 Todos los *items.*
* 0.20 Todos los *items.*

Esta gradación de los valores de c permitirá comprobar en qué medida se deterioran las estimaciones de los parámetros del Modelo de Rasch al aumentar c y en consecuencia la gravedad de la violación del supuesto c=0.

Valores de a:

* 1 Discriminación constante para todos los *items.*
* 0.50-2.00 Los valores de los Índices de Discriminación se encuentran entre 0.50 y 2.00 con Distribución Uniforme. Entre estos valores suelen encontrarse los hallados habitualmente, y nos permitirán comprobar la robustez de las estimaciones del Modelo de Rasch en presencia de discriminación variable de los *items* (a≠K).

Los seis valores de c cruzados con los de a (cxa) dan lugar a los 12 bloques de datos utilizados en la investigación y que se detallan a continuación:

BLOQUES DE DATOS SIMULADOS

Bloques	Valores de c y a	
n	c	a
1	0.00	1.00
2	0.00-0.50	1.00
3	0.50	1.00
4	0.33	1.00
5	0.25	1.00
6	0.20	1.00
7	0.00	0.50-2.00
8	0.00-0.50	0.50-2.00
9	0.50	0.50-2.00
10	0.33	0.50-2.00
11	0.25	0.50-2.00
12	0.20	0.50-2.00

VALORES DE OTROS PARÁMETROS

* Número de *Items:* 50
* Número de sujetos: 500
* θ: Los valores θ de los sujetos se generaron N(0,1), los mismos para los 12 bloques.

* b: Los Índices de Dificultad de los *items* se generaron con Distribución Uniforme entre -2 y $+2$, los mismos para los 12 bloques.

Siguiendo las recomendaciones de Lord (1975), los valores de los parámetros se eligieron de modo que se acercaran razonablemente a los valores empíricos que suelen tomar, para facilitar la generalización de los resultados a a situaciones reales. 50 *items* y 500 sujetos además de ser cifras habituales entre los usuarios de los modelos de TRI permiten obtener estimaciones ajustadas y estables de los parámetros. Parece poco aconsejable utilizar menos de 50 *items* cuando son de elección múltiple, especialmente si se utilizan fórmulas de corrección para controlar los efectos del azar.

Programa de simulación

Los datos se generaron con el programa DATAGEN (Hambleton y Rovinelli, 1973) en el centro de cálculo de la Universidad de Massachusetts (Amherst). El programa genera la matriz de aciertos y errores de los sujetos a los *items*. Elegidos los valores de los parámetros tal como se ha expuesto, el programa calcula la probabilidad $P(\theta)$ correspondiente a cada sujeto según la Curva Característica de parámetros conocidos. Posteriormente genera para cada sujeto un número al azar entre cero y uno. Si el número generado es igual o menor que la $P(\theta)$ del sujeto éste acierta el *item*, si es mayor lo falla. De este modo con un N suficiente la proporción de aciertos a determinado nivel tiende a $P(\theta)$.

Análisis de los datos

1. En primer lugar se estimaron los parámetros de los modelos logísticos de 1, 2 y 3 parámetros para cada bloque de datos simulados (36 análisis) mediante el programa LOGIST (Wingersky, Barton y Lord, 1982).

2. En segundo lugar se evaluó la precisión de las estimaciones anteriores por caminos concurrentes:

2.1. Correlaciones entre los valores simulados de los parámetros y los estimados.

2.2. Índice de Ajuste de θ: $IA(\theta)$.

$$IA(\theta) = \sqrt{\{\Sigma (\theta - \hat{\theta})^2\}/n}$$

donde θ son los valores simulados, $\hat{\theta}$ los estimados y n el número de sujetos. Nótese que aunque se simularon datos para 500 sujetos el programa LOGIST excluye de las estimaciones aquéllos que aciertan o fallan todos los *items*, por lo que no es infrecuente que se descarten algunos en cada análisis.

El índice $IA(\theta)$ complementa la información correlacional anterior (apar-

tado 2.1) ofreciendo una idea de la discrepancia entre los valores simulados y los estimados.

2.3. Bondad de Ajuste de las Curvas Características. Se utilizó el programa RESID (Hambleton, Murray y Simon, 1982) para obtener varios indicadores del Ajuste global de los modelos. El programa permite dividir la escala θ en 15 categorías (se usaron 12), calculando para cada una de ellas la diferencia (Residuo) entre los valores de $P(\theta)$ estimados y los simulados. La salida incluye datos detallados por categorías e *items* para varios índices: Residuos estandarizados, Residuos cuadráticos estandarizados, Residuos medios, Residuos medios absolutos y Residuos ponderados. Aquí se reseñarán (Tabla 1) las proporciones de Residuos Estandarizados (RE) con valores absolutos menores que 2. Si se asume la distribución Normal cabría esperar que el 95% de los Residuos tuvieran valores absolutos menores o iguales que 1.96, por lo que las proporciones ofrecidas constituyen un indicador intuitivo del ajuste de los modelos.

$$RE = \frac{P_e(\theta_j) - P_s(\theta_j)}{\sqrt{\{P_e(\theta_j)Q_e(\theta_j)/n\}}}$$

donde:

$P_e(\theta_j)$ es el valor correspondiente a la Curva Característica estimada para la categoría j.
$P_s(\theta_j)$ es el valor simulado para la misma categoría
$Q_e(\theta_j)$ es 1-Pe(θ_j)
n es el número de sujetos de la categoría j.

El grado de convergencia de los índices dará una idea bastante cabal de la Bondad de Ajuste, pero como señala Lord (1980), juicios con la solidez estadística que fuera de desear no son posibles dado el estado actual de conocimientos acerca de las distribuciones.

Resultados y discusión

Los resultados más destacables para el análisis de los objetivos propuestos aparecen sintetizados en la Tabla 1. En primer lugar se ofrecen las correlaciones entre los valores simulados de θ y los estimados para los modelos de uno (1-p), dos (2-p) y tres parámetros (3-p). A continuación aparecen las correlaciones entre los valores simulados de b y los estimados, luego el Índice de Ajuste de θ, y, finalmente, la proporción de residuos estandarizados con valores absolutos menores que 2.

Los resultados correspondientes a los 6 primeros bloques permiten comprobar la influencia de los valores del parámetro c (Aciertos al Azar) sobre la precisión de las estimaciones del Modelo de Rasch, así como evaluar si ésta varía al utilizar los Modelos Logísticos de 2 y 3 parámetros en vez del de Rasch.

TABLA 1. PRECISIÓN DE LAS ESTIMACIONES DE LOS PARÁMETROS DE LOS MODELOS LOGÍSTICOS PARA DISTINTOS VALORES DE LOS ACIERTOS AL AZAR (c) Y DE LOS ÍNDICES DE DISCRIMINACIÓN DE LOS *ITEMS* (a)

Bloques de datos			Correlaciones entre θ y $\hat{\theta}$: $r_{\theta\hat{\theta}}$			Correlaciones entre b y $\hat{b}$: $r_{b\hat{b}}$			Precisión de $\hat{\theta}$: IA($\hat{\theta}$)			Proporción de Residuos Estandarizados $<\|2\|$		
n	c	a	1-p	2-p	3-p	1-p	2-p	3-p	1-p	2-p	3-p	1-p	2-p	3-p
1	0.00	1.00	.973	.972	.968	.997	.997	.995	.238	.241	.258	.973	.980	.983
2	0.00-0.50	1.00	.920	.920	.935	.881	.793	.909	.401	.448	.368	.883	.965	.962
3	0.50	1.00	.838	.841	.873	.975	.748	.982	.615	.686	.525	.930	.982	.975
4	0.33	1.00	.889	.879	.914	.986	.933	.981	.484	.596	.445	.892	.970	.973
5	0.25	1.00	.920	.908	.936	.989	.953	.984	.402	.515	.467	.883	.963	.982
6	0.20	1.00	.929	.920	.935	.991	.955	.986	.375	.478	.375	.897	.962	.978
7	0.00	0.5-2.0	.975	.977	.973	.985	.997	.996	.224	.213	.235	.907	.977	.972
8	0.00-0.50	0.5-2.0	.918	.904	.935	.880	.794	.944	.454	.557	.392	.855	.950	.968
9	0.50	0.5-2.0	.846	.850	.878	.962	.706	.982	.634	.748	.561	.890	.965	.963
10	0.33	0.5-2.0	.893	.876	.929	.976	.908	.989	.534	.649	.396	.853	.953	.977
11	0.25	0.5-2.0	.921	.895	.940	.979	.939	.992	.423	.605	.376	.833	.948	.977
12	0.20	0.5-2.0	.935	.921	.947	.982	.946	.993	.372	.498	.345	.833	.947	.985

Como era de esperar (Tabla 1) la precisión de las estimaciones del modelo de Rasch decae ligeramente al incrementarse los Aciertos al Azar (c), pero curiosamente ocurre lo mismo con las estimaciones del modelo de 3 parámetros, lo que tal vez podría explicarse por la ya comentada imprecisión contaminadora de las estimaciones de c, constatada aquí con correlaciones entre los valores simulados y los estimados de 0.39 (bloque 2) y 0.604 (bloque 8) (Tabla 3). El ajuste general para el modelo de tres parámetros (Residuos) es algo mejor que el del modelo de Rasch, con proporciones superiores a 0.95 para los 6 bloques. Se confirman de este modo los datos de Van de Vijver (1986), las violaciones de $c=0$ parecen afectar más al ajuste del modelo como discrepancia simulado/estimado que a las correlaciones de las estimaciones. La precisión representada por IA(θ) también es ligeramente superior para el modelo de 3 parámetros. Desde un punto de vista estadístico ambos tipos de indicadores de ajuste (correlacionales y residuos) son equivalentes, pero el mensaje para los usuarios podría ser que si su interés fundamental se centra en el escalamiento de los sujetos en θ, o en el escalamiento de los *items* según su dificultad, casos habituales, las ordenaciones proporcionadas por el modelo de Rasch van a ser muy similares a las obtenidas con los modelos de 2 y 3 parámetros (véase además Tabla 2). Otro dato de interés aplicado se refiere a la cuantía de los desajustes (Residuos) en función de los valores de θ. Cuando $c\neq 0$ el ajuste es peor para valores bajos de θ, por lo que las estimaciones correspondientes a los sujetos con menor competencia en la variable medida serían menos fiables. El ajuste mejora notoriamente para valores medios y altos de θ. En situaciones de selección, por ejemplo, en las que el interés principal se

dirige a los sujetos superiores, sería irrelevante el modelo elegido. Este resultado era esperable, dado que al establecer $c \neq 0$ se impide que el valor asintótico de la Curva Característica sea 0. A la vista de los datos precedentes, no cabe sino confirmar la robustez del modelo de Rasch a violaciones de $c=0$, así como constatar la modesta mejora de precisión introducida, cuando lo hace, por el modelo de 3 parámetros. Nótese, por ejemplo, que de las cinco condiciones de violación (bloques 2 al 6), en tres de ellas (4,5,6) la correlación entre los valores de b estimados y los simulados es ligeramente superior para el modelo de Rasch que para el de 3 parámetros, teóricamente más indicado en tales circunstancias, lo cual no deja de ser sorprendente.

El segundo aspecto investigado fue la robustez de las estimaciones del modelo de Rasch cuando los *ítems* poseen diferentes Índices de Discriminación. Los datos básicos al respecto son los correspondientes al Bloque 7 en comparación con el Bloque 1 (Línea base: datos de 1 parámetro). De los cuatro índices de precisión ¡dos mejoran! [$r(\theta, \hat{\theta})$, $IA(\theta)$] y dos empeoran ligeramente ($r_{b\hat{b}}$, Residuos). La mejora de precisión, aunque mínima, en condiciones de violación es de difícil explicación, pero parece subrayar que las diferencias en general son exiguas y bien podría dar cuenta de ellas en la mayoría de los casos la mera variabilidad aleatoria. Las ganancias de precisión proporcionadas por los modelos de 2 y 3 parámetros que serían los adecuados en estas condiciones son más bien bajas, de nuevo algo mayores a nivel de residuos. Por ejemplo, el modelo de 3 parámetros genera ajustes peores que el de Rasch para $IA(\theta)$ y $r(\theta, \hat{\theta})$. De nuevo, tal vez la única explicación plausible sea el ruido introducido por la estimación imprecisa de c, lo que encaja con que aquí el modelo de 2 parámetros es algo más preciso que los de 1 y 3 en todos los casos, si bien las diferencias son exiguas. Se vuelven a confirmar por tanto los datos de Van de Vijver (1986): la presencia de índices de discriminación variables no parece afectar significativamente la precisión de las estimaciones del modelo de Rasch. Además, las ganancias de precisión aportadas por los modelos de 2 y 3 parámetros son mínimas, cuando son. Comparando los datos del bloque 7 con los de los bloques 2 al 6 puede comprobarse que las violaciones de discriminación variable afectan la precisión de las estimaciones incluso menos que las violaciones de $c=0$.

Finalmente, y en tercer lugar, la interacción de los valores de c y a (bloques 8 al 12) no aportan ninguna novedad relevante, confirmándose en términos generales la robustez de las estimaciones del modelo de Rasch, sobre todo a nivel correlacional, y la escasa ganancia en precisión (cuando se da) al utilizar los modelos de 2 y 3 parámetros en vez del modelo de Rasch. No parece que al darse ambas violaciones conjuntamente se produzca un efecto multiplicativo de la imprecisión de las estimaciones del modelo de Rasch. De nuevo el ajuste global (Residuos) sí es algo mejor en el modelo de 3 parámetros, con proporciones en los cinco bloques por encima de 0.95. Un caso típico sería el del bloque 8, en el que los *ítems* tienen valores c distintos de cero (entre 0.00 y 0.50) e Índices de Discriminación variable (entre 0.50 y 2.00), lo cual representa una situación habitual en la práctica para la mayoría de los usuarios de los modelos. Pues bien, ante violaciones tan flagrantes de los supuestos, el modelo de Rasch se desenvuelve más que aceptablemente (veáse Tabla 1, bloque 8). Las ganancias de precisión

generadas por el modelo de 3 parámetros (teóricamente el más indicado) frente al de Rasch no parece que vayan a convencer a muchos usuarios para cambiarse a un modelo más complejo: 0.017, 0.067, 0.062 y 0.113 respectivamente para $r(\theta, \hat{\theta})$, r_{bb}, $IA(\theta)$ y Residuos.

TABLA 2. CORRELACIONES ENTRE LAS ESTIMACIONES DE θ Y b POR LOS MODELOS LOGÍSTICOS DE 1, 2, Y 3 PARÁMETROS

Bloques			$r_{\hat{\theta}\hat{\theta}}$			r_{bb}		
n	c	a	$r_{\hat{\theta}1\,\hat{\theta}2}$	$r_{\hat{\theta}1\,\hat{\theta}3}$	$r_{\hat{\theta}2\,\hat{\theta}3}$	$r_{\hat{b}1\,\hat{b}2}$	$r_{\hat{b}1\,\hat{b}3}$	$r_{\hat{b}2\,\hat{b}3}$
1	0.00	1.00	.999	.996	.997	.998	.995	.998
2	0.00-0.50	1.00	.989	.986	.989	.947	.966	.960
3	0.50	1.00	.979	.975	.915	.797	.972	.733
4	0.33	1.00	.981	.945	.930	.914	.972	.967
5	0.25	1.00	.985	.966	.950	.927	.975	.977
6	0.20	1.00	.987	.968	.954	.934	.978	.976
7	0.00	0.50-2.00	.997	.993	.997	.986	.983	.999
8	0.00-0.50	0.50-2.00	.984	.969	.961	.935	.949	.918
9	0.50	0.50-2.00	.974	.885	.867	.735	.975	.723
10	0.33	0.50-2.00	.976	.929	.907	.874	.972	.923
11	0.25	0.50-2.00	.977	.950	.920	.906	.975	.944
12	0.20	0.50-2.00	.982	.961	.941	.920	.980	.854

Las escasas ventajas a nivel correlacional de los modelos de 2 y 3 parámetros frente al de Rasch pueden apreciarse observando en la Tabla 2 las elevadas correlaciones entre las estimaciones de los mismos parámetros (θ y b) por los tres modelos. Como ya se ha señalado, estos datos sugieren que si el interés del usuario se centra en el escalamiento de sujetos o *items* la elección de un modelo u otro va a tener una incidencia mínima en el ordenamiento.

Se confirma asimismo la imprecisión de las estimaciones de a y c (Tabla 3).

TABLA 3. CORRELACIONES ENTRE LOS VALORES SIMULADOS DE a Y c Y LOS ESTIMADOS POR LOS MODELOS DE 2 Y 3 PARÁMETROS

Bloques			$r_{c\hat{c}}$	$r_{a\hat{a}}$	
n	c	a	3-p	2-p	3-p
2	0.00-0.50	1.00	.390	——	——
7	0.00	0.50-2.00	——	.919	.860
8	0.00-0.50	0.50-2.00	.604	.361	.570
9	0.50	0.50-2.00	——	.372	.744
10	0.33	0.50-2.00	——	.373	.749
11	0.25	0.50-2.00	——	.392	.748
12	0.20	0.50-2.00	——	.418	.840

Las estimaciones de a (Modelo de 2 parámetros) resultan especialmente imprecisas cuando c≠0 y por tanto el modelo no se aplica en condiciones óptimas. Cuando c=0 la estimación es más precisa ($r_{a\hat{a}}$=0.919). Diríase que las estimaciones de a en el modelo de 2 parámetros son muy sensibles a la presencia de aciertos al azar. La estimación de c es más deficiente, con correlaciones de 0.390 y 0.604, confirmándose los datos de los autores ya citados.

Conclusiones generales

El objetivo fundamental de la investigación se centró en el estudio de la influencia que sobre las estimaciones del modelo de Rasch tiene el incumplimiento de dos condiciones exigidas por el modelo, a saber, la inexistencia de aciertos al azar y la discriminación constante de los *items*. La importancia de conocer el comportamiento del modelo bajo estas circunstancias tiene por supuesto gran interés teórico-estadístico, pero lo tiene sobre todo aplicado, dado el uso masivo del modelo en situaciones en las que existe una alta probabilidad, rayana con la certeza, de que no se cumplan los citados supuestos. No sólo se trataba de conocer el comportamiento del modelo de Rasch en condiciones adversas, sino de evaluar sistemáticamente las posibles ganancias de precisión aportadas, si tal hubiere, por los modelos logísticos de 2 y 3 parámetros, teóricamente más apropiados en tales circunstancias. Los resultados detallados y comentados previamente confirman que las estimaciones del modelo de Rasch, y por ende su ajuste, se ven poco afectadas por la presencia de Aciertos al Azar (c≠0) y menos aún por la de Índices de Discriminación variable ($a_i \neq K$). Por su parte, las ganancias en la precisión de las estimaciones proporcionadas por los modelos de 2 y 3 parámetros frente al de Rasch son ciertamente escasas. Incluso en los casos en los que tales diferencias existen, su cuantía es exigua, por lo que difícilmente justifica el uso de modelos más complejos y que además tienen ciertos problemas en la estimación de algunos de sus parámetros, constatándose efectivamente la imprecisión de las estimaciones de a y c, especialmente c. De todo lo cual se colige que el uso del modelo de Rasch en circunstancias similares a las aquí simuladas no parece descabellado. Obviamente, ello no quiere decir que las situaciones reales respondan estrictamente, ni mucho menos, a la asepsia de los datos simulados, pero constituye seguramente la mejor modelización posible. Por ejemplo, aquí se asume la unidimensionalidad perfecta, rara avis en la práctica y que afecta las estimaciones. Además de la unidimensionalidad, no conviene olvidar la incidencia del número de *items* y sujetos, así como de los valores y distribuciones de θ y b. Se utilizaron valores que reflejasen situaciones reales probables, pero debido, entre otros, a los factores citados no controlados, las generalizaciones han de ser necesariamente tentativas. Sólo esperar que, al igual que la naturaleza imita al arte...., los datos reales lo hagan con los simulados.

RESUMEN

Se investigó la influencia que tienen sobre la precisión de las estimaciones del Modelo de Rasch la violación del supuesto de Inexistencia de Aciertos al Azar (c=0) y de Discriminación constante de los *Items* (a=K). Asimismo se evaluó la ganancia de precisión respecto del Modelo de Rasch generada por los modelos Logísticos de 2 y 3 parámetros, teóricamente más apropiados en esas circunstancias. Se utilizaron 12 bloques de datos simulados provenientes de cruzar 6 valores de c con 2 de a. El número de *items* fue 50, el de sujetos 500, los valores de θ se distribuyeron N(0,1) y los de b con distribución uniforme entre -2 y +2. Los datos se simularon con el programa DATAGEN y los parámetros se estimaron con LOGIST. La Bondad de Ajuste de los modelos se evaluó mediante el análisis de residuos proporcionado por el programa RESID y las correlaciones entre los valores simulados y estimados de los parámetros; en el caso de θ se utilizó además otro índice adicional. A la vista de los resultados puede concluirse que las estimaciones del Modelo de Rasch son aceptablemente robustas a la presencia de Aciertos al Azar (c≠0) y altamente robustas a Índices de Discriminación variables (a≠K). Por su parte, los modelos Logísticos de 2 y 3 parámetros no introducen mejoras de precisión significativas en relación con el modelo de Rasch.

SUMMARY

The robustness of Rasch model estimates to violations of assumptions, zero Guessing and equal Item Discriminations, were studied. The gains in the estimates accuracy when using the 2 and 3 parameter Logistic models instead of Rasch model were also evaluated. Twelve simulated data sets were generated by crossing 6 values of c with two values of a. The number of items was 50, the number of subjects 500, the values of θ were N(0,1) and b values were Uniformly distributed between -2 and +2. LOGIST was used to estimate the parameters and several indices were calculated to evaluate model fit. The results seem to indicate that Rasch model estimates are acceptably robust to Guessing and highly robust to heterogeneity in Item Discrimination. At the same time, the gains in accuracy when using the 2 and 3 parameter Logistic models instead of the Rasch model were modest.

REFERENCIAS

Andrich, D. (1988). *Rasch models for measurement.* Beverly Hills, California: Sage Publications.

Ansley, T.N., y Forsyth, R.A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two dimensional data. *Applied Psychological Measurement, 9,* 37-48.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En F.M. Lord y M. Novick, *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley Publishing Company.
Bock, R.D., y Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.
Dinero, T.E., y Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations, *Applied Psychological Measurement, 4*, 581-592.
Divgi, D.R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement, 23*, 283-298.
Drasgow, F., y Parsons, C.K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199.
Greaud, V.A. (1988). Some effects of applying unidimensional IRT to multidimensional tests. Paper presented at the AERA annual meeting, New Orleans.
Hambleton, R.K. (Ed.) 1983. *Applications of item response theory.* Vancouver, BC: Educational Research Institute of British Columbia.
Hambleton, R.K., y Cook, L.L. (1983). Robustness of item, response models and effects of test length and sample size on the precision of ability estimates. En D.J. Weiss (Ed.), *New horizons in testing.* New York: Academic Press.
Hambleton, R.K., Murray, L.N., y Simon, R. (1982). Utilization of item response models with NAEP mathematics exercise results. Final Report (NIE-ECS Contract No. 02-81-20319). Washington, DC: National Institute of Education.
Hambleton, R.K., y Rovinelli, R.A. (1973). A FORTRAN IV program for generating examinee response data from logistic test models. *Behavioral Science, 17*, 73-74.
Hambleton, R.K., y Swaminathan, H. (1985). *Item response theory:* principles and applications. Boston, MA: Kluwer-Nijhoff.
Hambleton, R.K., y Traub, R.E. (1971). Information curves and efficiency of three logistic test models. *British Journal of Mathematical and Statistical Psychology, 24*, 372-381.
Harrison, D.A. (1986). Robustness of IRT parameter estimation to violation of the unidimensionality assumption. *Journal of Education Statistics, 11*, 91-115.
Hattie, J.A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.
Hulin, C.L., Drasgow, F. y Parsons, C.K. (1983). *Item, response theory: Application to psychology measurement.* Hornewood, III: Dow Jones-Irvin.
Hulin, C.L., Lissak, R.I. y Drasgow, F. (1982). recovery of two and three parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.
Kolen, J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18*, 1-11.
Lord, F.M. (1975). Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. *Research Bulletin 75-33.* Princeton, NJ: Educational Testing Service.
Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: LEA.
Lord, F.M. (1983). Small N justifies Rasch model. En D.J. Weiss (Ed.), *New horizons in testing.* New York: Academic Press.
Lord, F.M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23(2)*, 157-162.
Lord, F.M., y Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley Publishing Company.
Lumsden, J. (1961). The construction of unidimensional tests. *Psychological Bulletin, 58*, 122-131.
Lumsden, J. (1976). Test theory. *Annual Review of Psychology, 27*, 251-280.
McKinley, R.L., y Mills, C.N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*, 49-57.
Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-196.
Mislevy, R.J., y Bock, R.D. (1984). *BILOG Version 2.2: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhague: The Danish Institute for Educational Research.
Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4(3)*, 207-230.
Ree, J.M. (1979). Estimating item characteristic curves. *Applied Psychological Measurement, 3*, 371-385.
Rentz, R.R., y Rentz, C.C. (1978). *Does the Rasch model really work?* ERIC Report No. 67. Princeton, NJ.
Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39*, 111-121.

Swaminathan, H., y Gifford, J.A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7*, 175-192.
Swaminathan, H., y Gifford, J.A. (1983). Estimation of parameters in the three-parameter latent trait model. En D.J. Weiss (Ed.), *New horizons in testing*. New York: Academic Press.
Swaminathan, H., y Gifford, J.A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50*, 349-364.
Swaminathan, H., y Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51*, 589-601.
Thissen, D., y Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501-519.
Thissen, D., y Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-577.
Thissen, D., y Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*, 397-412.
Traub, R.E. (1983). A priori considerations in choosing an item response model. En R.K. Hambleton (Ed.), *Aplications of item response theory*, Vancouver, BC: Educational Research Institute of British Columbia.
van de Vijver, F.J. (1986). The robutsness of Rasch estimates. *Applied Psychological Measurement, 10*, 45-57.
Wainer, H., y Wright, B.D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika, 45*, 373-391.
Whitely, S. (1988). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479-494.
Wingersky, M.S., Barton, M.A. y Lord, F.M. (1982). *LOGIST 5.0 Version 1.0 User's Guide*. Princeton, NJ: Educational Testing Service.
Wingersky, M.S., y Lord, F.M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8*, 347-364.
van den Wollenberg, A.L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-140.
Wright, B.D. (1977a). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.
Wright, B.D. (1977b). Misunderstanding of the Rasch model. *Journal of Educational Measurement, 14*, 219-226.
Wright, B.D., Mead, R.J. y Bell, S.R. (1979). *BICAL: A Rasch Program for the Analysis of Dichotomous Data*. Chicago: MESA Press.
Wright, B.D., y Stone, M.H. (1979). *Best Test Design*. Chicago: MESA Press.
Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.
Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.
Yen, W.M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*, 275-291.