Psicothema 2009. Vol. 21, nº 4, pp. 639-645
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Bondad de ajuste en ítems politómicos: tasas de error tipo I y potencia de tres índices de ajuste

Manuel J. Sueiro y Francisco José Abad*
Universidad Complutense de Madrid y * Universidad Autónoma de Madrid

Al aplicar un modelo de Teoría de la Respuesta al Ítem es fundamental disponer de un procedimiento que permita conocer si el modelo se ajusta a los datos. Este artículo compara, mediante un estudio de simulación, las tasas de error tipo I y potencia de tres tipos de índices de ajuste generalizados a ítems politómicos: el índice tradicional basado en la agrupación de los sujetos según su nivel de rasgo estimado, otro basado en el cálculo de las probabilidades posteriores y un tercero consistente en agrupar a los sujetos mediante su puntuación total en el test. Las condiciones bajo estudio fueron la longitud del test (10, 20 y 40 ítems), número de opciones de los ítems (3, 4 y 5) y tamaño de la muestra (500, 1.000 y 2.000 sujetos). Los resultados mostraron que el índice basado en las probabilidades posteriores presentaba tasas de error más próximas a las nominales, así como una mayor potencia, especialmente cuando la muestra era grande o el test era corto.

Goodness of fit in polytomous items: Type I error rates and empirical power for three fit indexes. Applications of Item Response Theory require assessing the agreement between observations and model predictions at the item level. This paper compares approaches applied to polytomous scored items in a simulation study. Three fit-indexes are calculated: traditional chi-square index obtained by grouping examinees according to their estimated trait, an alternative that uses posterior distribution of trait and the third method, in which examinees are grouped according their observed total scores. Various conditions are simulated by manipulating test length (10, 20 and 40 items), number of categories (3, 4 and 5) and sample size (500, 1000 and 2000 examinees). Power and Type I error rates are described. Chi-square statistics based on posterior probabilities showed the best performance, especially with larger sample sizes and shorter test lengths.

La Teoría de la Respuesta al Ítem (TRI) permite modelar matemáticamente las relaciones entre el nivel de rasgo de los sujetos y la probabilidad de cada respuesta a un ítem. La TRI permite el desarrollo de aplicaciones psicométricas como los tests adaptativos informatizados o la equiparación de tests, en las que aunque las personas respondan a ítems distintos podemos estimar los niveles de rasgo en la misma escala métrica (Olea y Ponsoda y Prieto, 1999). Sin embargo, la utilidad de los modelos depende del grado de correspondencia entre lo que predice el modelo y los datos empíricos. El estudio del ajuste individual de cada ítem permite establecer qué elementos ofrecen un mal ajuste y deben ser descartados o rediseñados. Se han propuesto diversos procedimientos para evaluar la bondad de ajuste de un modelo de TRI (Swaminathan, Hambleton y Rogers, 2007), pero presentan algunas dificultades como se describirá a continuación.

La estrategia general consiste en comparar las frecuencias de respuesta predichas por el modelo y las observadas para un ítem *j*, de dos maneras alternativas:

Fecha recepción: 13-11-08 • Fecha aceptación: 11-3-09
Correspondencia: Manuel J. Sueiro
Facultad de Psicología
Universidad Complutense de Madrid
28223 Madrid (Spain)
E-mail: msueiro@psi.ucm.es

– Como un estadístico de bondad de ajuste de Pearson:

$$\chi^2_j = \sum_Q \sum_K \frac{n_q\left(O_{qk} - E_{qk}\right)^2}{E_{qk}}$$

– Como prueba de razón de verosimilitud:

$$G^2_j = 2\sum_Q \sum_K n_q O_{qk} \ln\left(\frac{O_{qk}}{E_{qk}}\right)$$

Siendo Q el número de subgrupos, homogéneos en el nivel de rasgo, en los que se ha clasificado la muestra, K el número de alternativas de respuesta del ítem y n_q la frecuencia absoluta de sujetos en el grupo q; O_{qk} y E_{qk} son, respectivamente, las proporciones observada y esperada de personas que escogen la alternativa de respuesta k en el subgrupo q. Tradicionalmente se ha asumido que ambos estadísticos siguen una distribución χ^2 con $Q^*(K-1) - m$ grados de libertad, siendo m el número de parámetros estimados. Sin embargo, al ser el rasgo una variable latente y la clasificación una discretización arbitraria de una variable continua, la distribución real de los estadísticos es desconocida.

En la aproximación tradicional se forman los Q grupos (por ejemplo, 10 grupos) y se comparan las probabilidades observadas

y esperadas en función de los niveles de rasgo estimados $\ddot{\theta}$ (Bock, 1972; Yen, 1981; McKinley y Mills, 1985). Estos estadísticos presentan tasas de error tipo I inaceptables cuando el nivel de rasgo no se estima con suficiente precisión. Además, tanto el número de intervalos empleados como el criterio utilizado para el agrupamiento pueden afectar a su funcionamiento y tienen un carácter arbitrario (Reise, 1990). Por ello, en los últimos años, se han propuesto algunos índices que pretenden solucionar estos problemas.

Índices basados en las puntuaciones totales de los sujetos. Orlando y Thissen (2000) proponen formar los Q grupos y comparar las probabilidades observadas y esperadas en función de las puntuaciones en el test (s), formando inicialmente tantos grupos como puntuaciones. En cada grupo de sujetos de igual puntuación en el test podemos observar la proporción de sujetos que acierta cada ítem, con la ventaja sobre la forma de la distribución de que la puntuación total es un dato observable y no latente. La probabilidad esperada de elegir la opción k para los que tienen la puntuación s se define como:

$$P\left(x_{ij}=1\middle|S_i=s\right)=\frac{\int P\left(x_{ij}=k\middle|\theta\right)P\left(S_i^j=s-1\middle|\theta\right)g(\theta)d\theta}{\int P\left(S_i=s\middle|\theta\right)g(\theta)d\theta}$$

Donde $P(S_i^j = s-1 \mid \theta)$ indica la probabilidad de obtener la puntuación *s-1* en el test formado por todos los ítems excepto el ítem *j*.

El problema de este índice reside en su complejidad computacional: el cálculo de las probabilidades esperadas requiere el uso del algoritmo iterativo de Lord y Wingersky (1984; desarrollado en Thissen, Pommerich, Billeaud y Williams, 1995) y puede emplear bastante tiempo si el número de ítems es alto. Esto es especialmente relevante en la generalización del índice a modelos politómicos. Además, el incremento en el número de celdillas dejará muchas casillas vacías, lo que puede afectar a la distribución del estadístico y exige un procedimiento para tratar con ellas.

Índices basados en las probabilidades posteriores. Stone (2000) propone modificar el modo en que se calculan las frecuencias observadas. En lugar de emplear el rasgo estimado del sujeto se emplea su distribución posterior, $P(\theta = \theta_q \mid \mathbf{X}_i = \mathbf{x})$. Ésta indica la probabilidad de que el nivel de rasgo esté comprendido en cada subgrupo q del continuo supuesto su patrón de respuestas $\boldsymbol{x}$. Cuanto mayor sea la imprecisión de la estimación del rasgo, más distribuidas en los subgrupos estarán las probabilidades del sujeto. Por ejemplo, sumando las probabilidades posteriores a través de todos los sujetos que escogen una opción k se obtendrán las *pseudo-frecuencias observadas*:

$$n_{qjk}^{*}=\sum_i P\left(\theta=\theta_q\middle|\,\mathbf{X_i}=\mathbf{x};x_{ij}=k\right)$$

Un problema de esta aproximación es que las probabilidades posteriores subyacentes a la distribución de pseudo-frecuencias no son independientes, por lo que no puede asumirse la distribución χ^2. Stone (2000) propone factores de corrección, obtenibles mediante *bootstrap*, que permiten reescalar los índices y sus grados de libertad, mostrando que dichos factores de corrección aproximan la distribución a χ^2.

Comparación de los distintos tipos de índices de ajuste. El número de ítems y el tamaño muestral han sido los factores más estudiados. El número de ítems es determinante a la hora de establecer la precisión de la estimación del rasgo, utilizado para agrupar a los sujetos cuando se calcula el índice tradicional. Los efectos del tamaño muestral son más complejos. Por un lado, la existencia de casillas con frecuencias muy bajas afecta negativamente a la distribución de estos índices. Este problema es mayor cuanto mayor es el número de grupos que se forman. Por otro lado, cuanto mayor es el tamaño muestral, mayor es la potencia para detectar el verdadero desajuste, pero también para detectar las discrepancias producidas por la incorrecta agrupación de los sujetos dado $\hat{\theta}$.

Orlando y Thissen (2000, 2003) comparan el rendimiento del índice tradicional de Yen (1981) con su propuesta basada en las puntuaciones totales en ítems dicotómicos. Encontraron un incorrecto funcionamiento del índice tradicional cuando el test era corto (i.e., 10 ítems) siendo inferior al funcionamiento de su propuesta, que mostraba tasas de error cercanas al valor nominal. El funcionamiento de ambos indicadores mejora con la longitud del test (i.e., 40 ítems), pero en presencia de tests muy largos (i.e., 80 ítems) y bajos tamaños muestrales (i.e., 500 sujetos) el nuevo índice puede llegar a mostrar menor potencia, probablemente por el problema de las casillas vacías.

Stone y Zhang (2003), también con dicotómicos, compararon los tres tipos de índices con similares resultados. El índice basado en las probabilidades posteriores mostró tasas de error próximas a las nominales y los mejores resultados en cuanto a la potencia, sobre todo en muestras pequeñas.

Por su parte, Glas y Suárez-Falcon (2003) encontraron que el índice basado en la puntuación total mostraba mejor funcionamiento que el tradicional. Sin embargo, para ambos índices, en presencia de ítems desajustados (i.e., 10%) se incrementaba la tasa de falsas alarmas especialmente si la muestra era muy grande (i.e., 4.000 sujetos). Este efecto se producía incluso si el test era largo (i.e., 40 ítems).

La mayor parte de la investigación se ha centrado en los ítems dicotómicos. En ítems politómicos se ha visto que los índices G^2 tradicionales incluidos en programas como PARSCALE (Muraki y Bock, 1997) han mostrado un funcionamiento inadecuado (DeMars, 2005). Recientemente, algunos autores (Kang y Chen, 2007; Roberts, 2008) han propuesto generalizaciones del índice de Orlando y Thissen a ítems politómicos. Aunque en estos estudios el índice de Orlando y Thissen muestra un mejor funcionamiento que el indicador tradicional, no se ha comparado su eficacia en comparación con otros índices como el propuesto por Stone. Esto es importante puesto que se espera que el problema de las casillas vacías sea mayor para los índices basados en la puntuación, pues en ítems politómicos aumenta el número de puntuaciones posibles y se puede incrementar el número de casillas con frecuencias muy bajas o nulas (Stone y Zhang, 2003).

En el presente trabajo se compara el uso de los distintos indicadores para uno de los modelos politómicos más usuales, el modelo de respuesta graduada (Samejima, 1969), y se analiza el efecto sistemático del número de opciones, el tamaño de la muestra, la longitud del test y la presencia de ítems desajustados en el funcionamiento de estos indicadores.

Método

Parámetros de los ítems. Los parámetros de los ítems se construyeron para formar un test «completo y conveniente» (Ankenmann, Witt y Dunbar, 1999). Se partió de la formulación que hace Muraki (1990) del modelo de respuesta de Samejima:

$$P*\left(x_{ij} \geq k \middle| \theta = \theta_i\right) = \frac{1}{1+\exp\left[-a_j\left(\theta_i - b_j + c_{k-1}\right)\right]}$$

Donde el parámetro b_{jk} del modelo de respuesta graduada se descompone en dos parámetros: un parámetro b_j de localización distinto para cada ítem *j* y un parámetro de categoría c_k distinto para cada categoría *k*, siendo el conjunto de parámetros *c* iguales para todos los ítems. Se buscó un conjunto de parámetros *c* que cubrieran todo el rango del rasgo (i.e., entre -2 y 2) dividiéndolo en zonas de igual tamaño para las condiciones de 3, 4 y 5 alternativas de respuesta. Los parámetros de localización b_j se escogieron para que las b_{jk} finales tuvieran media y desviación típica, próximas a 0 y 1, respectivamente. Respecto al parámetro de discriminación *a* se utilizaron dos niveles: 1.2 y 2.2.

Combinando estos parámetros se obtuvieron tres conjuntos de diez ítems para las condiciones de 3, 4 y 5 alternativas, tal y como se muestra en la tabla 1. Estos ítems se replicaron para formar los tests de 20 y 40 ítems.

Factores bajo estudio. Se incluyeron como factores, el tamaño muestral (500, 1.000 y 2.000 sujetos), la longitud del test (10, 20 y 40 ítems) y el número de alternativas de respuesta (3, 4 y 5). Además se consideró una condición adicional: ausencia o presencia de ítems desajustados en el test, para comparar las tasas de error tipo I que se producen cuando todos los ítems ajustan con las que se producen cuando algunos están desajustados y se incrementa el error al clasificar a los sujetos. El desajuste se introdujo en el *33%* de los ítems de cada condición, de la siguiente manera:

1. En el parámetro a_j, en los ítems 8, 18, 28 y 38 (dependiendo de la longitud del test). Este desajuste se provocó restando –0.5 al parámetro *a* estimado.
2. En los parámetros b_{jk}, en los ítems 3, 13, 23 y 33 (siguiendo la misma lógica del caso anterior), restando –0.25 a todos los parámetros *b* del ítem estimado.
3. Tanto en *a* como en los parámetros *b* estimados, en los ítems 9, 19, 29 y 39.

Generación de las respuestas simuladas. Los datos fueron simulados utilizando *MRFITIT* (Sueiro y Abad, en preparación) del siguiente modo: primero se generaron aleatoriamente los parámetros de los sujetos según una distribución *N*(0,1). A partir de ellos se generaron las respuestas calculando la probabilidad acumulada de cada alternativa de respuesta y comparándola con un valor aleatorio según una distribución uniforme (0,1). Para cada una de las 54 condiciones experimentales se obtuvieron 100 réplicas.

Calculando los estadísticos de ajuste. Para cada réplica se reestimaron los parámetros de los ítems a partir de las respuestas simuladas utilizando MULTILOG 7.0, fijando en 100 el número máximo de iteraciones para garantizar la convergencia y dejando el resto de las opciones por defecto (Thissen, Chen y Bock, 2003). Fueron estas estimaciones las que se utilizaron en las pruebas de bondad de ajuste. Para calcular los índices de ajuste se utilizó *MRFITIT* (Sueiro y Abad, en preparación). El índice tradicional se calculó en base a los deciles del nivel de rasgo estimado, utilizando la media de cada grupo para calcular las probabilidades esperadas. Ya que la prueba de χ^2 es tremendamente sensible a las casillas con valores muy bajos se colapsaron las celdillas con una frecuencia esperada igual o menor que 2, colapsando primero intragrupo las categorías de respuesta y, cuando esto no solucionara el problema, colapsando los grupos adyacentes. La significación de este índice se obtuvo asumiendo su distribución χ^2.

El índice basado en las probabilidades posteriores se calculó según las especificaciones de Stone (2000). Para reescalar el estadístico se realizó *bootstrap* con 100 muestras para calcular las constantes de escalamiento (siguiendo lo descrito por Stone, 2000). Su significación estadística se evaluó asumiendo dicha distribución χ^2 escalada.

Para el índice basado en las puntuaciones totales observadas se siguió una aproximación análoga a la del cálculo del índice tradicional (utilizando la puntuación total observada en lugar del la θ estimada para formar los grupos), colapsando de igual modo. De igual manera, la significación estadística del índice se obtuvo asumiendo su distribución χ^2.

Los índices fueron calculados en su forma de estadístico de bondad de ajuste de Pearson y en la de prueba de razón de verosimilitud.

Resultados

La tabla 2 muestra las tasas de error Tipo I obtenidas en la situación de ajuste en cada condición.

Tabla 1
Parámetros verdaderos de los ítems simulados

Ítems de 3 alternativas				**Ítems de 4 alternativas**					**Ítems de 5 alternativas**					
Item	***a***	***b1***	***b2***	***Item***	***a***	***b1***	***b2***	***b3***	***Item***	***a***	***b1***	***b2***	***b3***	***b4***
1	1.2	-1.77	-.43	1	1.2	-1.8	-.8	.2	1	1.2	-1.8	-1.0	-.2	.6
2	1.2	-1.07	.27	2	1.2	-1.3	-.3	.7	2	1.2	-1.4	-.6	.2	1.0
3	1.2	-.67	.67	3	1.2	-1.0	.0	1.0	3	1.2	-1.2	-.4	.4	1.2
4	1.2	-.27	1.07	4	1.2	-.7	.3	1.3	4	1.2	-1.0	-.2	.6	1.4
5	1.2	.43	1.77	5	1.2	-.2	.8	1.8	5	1.2	-.6	.2	1.0	1.8
6	2.2	-1.77	-.43	6	2.2	-1.8	-.8	.2	6	2.2	-1.8	-1.0	-.2	.6
7	2.2	-1.07	.27	7	2.2	-1.3	-.3	.7	7	2.2	-1.4	-.6	.2	1.0
8	2.2	-.67	.67	8	2.2	-1.0	.0	1.0	8	2.2	-1.2	-.4	.4	1.2
9	2.2	-.27	1.07	9	2.2	-.7	.3	1.3	9	2.2	-1.0	-.2	.6	1.4
10	2.2	.43	1.77	10	2.2	-.2	.8	1.8	10	2.2	-.6	.2	1.0	1.8

Tabla 2
Tasas de error empíricas para los distintos índices de ajuste en la situación de todos los ítems ajustados (α= .05)

Número de alternativas, longitud del test (ítems) y tamaño de la muestra (N)			Métodos					
			Stone (2000)		Yen (1981), McKinley y Mills (1985)		Orlando y Thissen (2003)	
Alternativas	N	ítems	G^2*	X^2*	Q_1-G^2	Q_1-X^2	S-G^2	S-X^2
3	500	10	0.02	0.03	0.49	0.33	0.06	0.05
		20	0.02	0.03	0.10	0.06	0.06	0.04
		40	0.03	0.04	0.07	0.05	0.06	0.04
	1000	10	0.02	0.02	0.67	0.62	0.06	0.05
		20	0.02	0.03	0.19	0.11	0.07	0.06
		40	0.03	0.04	0.07	0.05	0.05	0.05
	2000	10	0.01	0.01	0.93	0.89	0.06	0.05
		20	0.02	0.02	0.38	0.27	0.05	0.04
		40	0.03	0.04	0.09	0.07	0.05	0.05
4	500	10	0.01	0.02	0.34	0.18	0.07	0.05
		20	0.02	0.04	0.10	0.05	0.07	0.05
		40	0.03	0.05	0.07	0.05	0.07	0.05
	1000	10	0.02	0.03	0.61	0.50	0.07	0.06
		20	0.02	0.04	0.14	0.07	0.06	0.05
		40	0.03	0.04	0.07	0.05	0.07	0.05
	2000	10	0.01	0.01	0.73	0.66	0.05	0.05
		20	0.02	0.02	0.23	0.14	0.06	0.06
		40	0.03	0.04	0.08	0.06	0.06	0.05
5	500	10	0.01	0.03	0.30	0.13	0.08	0.05
		20	0.02	0.04	0.11	0.05	0.06	0.04
		40	0.03	0.05	0.08	0.05	0.07	0.04
	1000	10	0.01	0.01	0.50	0.31	0.05	0.04
		20	0.03	0.04	0.11	0.06	0.06	0.05
		40	0.03	0.04	0.07	0.05	0.07	0.05
	2000	10	0.01	0.02	0.64	0.59	0.05	0.04
		20	0.03	0.04	0.19	0.11	0.06	0.05
		40	0.04	0.04	0.08	0.06	0.06	0.05

Los índices tradicionales (Q_1-X^2 y Q_1-G^2) presentaron tasas de error superiores al nivel nominal (α= 0.05) prácticamente en todos los casos. Su rendimiento fue peor en la condición de 10 ítems y en la condición de 2.000 sujetos. Sólo cuando el test era largo (40 ítems) mostraron un funcionamiento razonable, especialmente si la muestra no era grande. Por su parte, los índices basados en las probabilidades posteriores mostraron un desempeño notablemente mejor, si bien resultaron excesivamente conservadores, con tasas de error tipo I inferiores al nivel de significación .05, especialmente en la forma de razón de verosimilitud. En lo que respecta a los índices basados en la puntuación total, mostraron también un rendimiento bastante adecuado, si bien las tasas de error fueron ligeramente mayores a lo esperado en el índice de razón de verosimilitud S-G^2.

Se observa una relación entre la longitud del test y el rendimiento del índice tradicional: mejor cuanto más largo, congruente con la idea de que es la imprecisión en la estimación de los rasgos latentes de los sujetos la que hace inutilizable este índice. Así, cuanto mayor es el tamaño muestral, mayor es la potencia para detectar el desajuste provocado por una mala agrupación de los sujetos según el nivel de rasgo (mal) estimado y peor es el rendimiento de estos índices.

La tabla 3 resume las tasas de errores de detección incorrectos (falsas alarmas) obtenidas en la situación de desajuste, es decir, la proporción de ítems que fueron erróneamente señalados como desajustados del total de ítems que estaban realmente ajustados. Esto informa de las tasas de error tipo I de los distintos índices en una situación en la que no todos los ítems ajustaban al modelo. De nuevo los índices basados en las estimaciones del rasgo latente presentan resultados inaceptables, siendo éstos función de la longitud del test (mejor rendimiento en tests largos) y del tamaño muestral (peor rendimiento en muestras grandes).

En lo que respecta a los nuevos índices, en cualquiera de sus formas, su funcionamiento es notablemente mejor que el de los índices tradicionales. En este caso se obtienen resultados más próximos al nivel de significación, ligeramente superiores en algunos casos especialmente para el índice de Orlando y Thissen, lo que parece indicar que en presencia de ítems desajustados los nuevos índices pierden alguna capacidad discriminativa. En ambos índices se observa un peor rendimiento cuando el tamaño muestral aumenta. Para el índice de Stone el problema tiende a reducirse cuando aumenta el número de alternativas.

La tabla 4 muestra las tasas de detecciones correctas (potencia) para los ítems que presentaban desajuste.

Los valores de potencia en el índice tradicional no son, en general, interpretables, teniendo en cuenta las elevadas tasas de error tipo I obtenidas. Podemos hablar de falta de especificidad del índi-

ce para detectar el desajuste. Con el índice de Stone se obtuvieron los mejores resultados: la potencia alcanzó el 100% en 16 de las 27 condiciones y obtuvo valores superiores al 60% en el resto. Su funcionamiento fue mejor cuanto mayor fue el número de sujetos y cuanto menores la longitud del test y el número de alternativas. Con el índice de Orlando y Thissen se obtuvieron peores resultados. La tasa de detección fue baja en muestras pequeñas. Aunque aumentaba con el tamaño muestral, la tasa de falsas alarmas se incrementaba también, lo que hace que los valores de las tasas de potencia no sean interpretables. La longitud del test y el número de alternativas no parecieron tener ningún efecto en este caso.

Discusión y conclusiones

El problema del ajuste de los modelos es fundamental en la TRI y el desarrollo de índices fiables que permitan decidir si el grado de ajuste entre los datos y el modelo es adecuado ha generado abundante investigación. Este estudio compara algunos de los procedimientos de ajuste que se han propuesto recientemente y evalúa su eficacia en ítems politómicos y bajo diferentes condiciones. Los resultados muestran que los índices tradicionales basados en la habilidad estimada de los sujetos (como el de Bock, 1972; Yen, 1981) son ineficaces para detectar ítems desajustados. Estos índices presentan tasas de error tipo I extremadamente elevadas, lo que indica realmente falta de especificidad, especialmente en tests cortos o con tamaños muestrales elevados. Ya que en la estimación de parámetros en la TRI ambas variables están estrechamente relacionadas (son necesarias muestras más grandes para estimar los parámetros de tests más largos) se presenta un escollo difícil de superar para estos índices.

De las alternativas propuestas, sobresale la de Stone (2000), cuyo índice basado en las probabilidades posteriores presenta un rendimiento muy adecuado, tanto en sus tasas de error como en su capacidad para detectar el desajuste, siendo bastante robusto y aparentemente poco dependiente de las condiciones (tamaño muestral, número de alternativas y longitud del test) en las que se utiliza. Sin embargo, el índice de Stone presenta el problema en absoluto trivial de no tener una distribución conocida. Su aplicación exige la construcción de su distribución empírica simulándola mediante *bootstrapping*. Esta solución hace pensar si el éxito del índice no se basará en dicha construcción simulada de su distribución, ya que resultados bastante prometedores al respecto se han encontrado también utilizando índices tradicionales cuando en lugar de asumir su distribución χ^2 se ha recurrido al *bootstrapping* para interpretar su significación estadística (von Davier, 1997).

Los índices de Orlando y Thissen (2000), si bien más elegantes que el de Stone y con un rendimiento semejante aunque ligeramente inferior, no están exentos de problemas. Su aplicación en

Tabla 3
Falsas alarmas (detecciones de desajuste incorrectas) para los distintos índices de ajuste en la condición de desajuste en algunos ítems (α= .05)

Número de alternativas, longitud del test (ítems) y tamaño de la muestra (N)			Métodos					
			Stone (2000)		Yen (1981), McKinley y Mills (1985)		Orlando y Thissen (2003)	
Alternativas	N	ítems	G^2*	X^2*	Q_1-G^2	Q_1-X^2	S-G^2	S-X^2
3	500	10	0.03	0.04	0.54	0.40	0.09	0.08
		20	0.04	0.05	0.12	0.07	0.09	0.08
		40	0.05	0.07	0.09	0.07	0.09	0.07
	1000	10	0.07	0.08	0.70	0.61	0.09	0.08
		20	0.06	0.07	0.24	0.15	0.12	0.12
		40	0.07	0.08	0.10	0.08	0.10	0.09
	2000	10	0.20	0.21	0.97	0.94	0.19	0.17
		20	0.16	0.16	0.46	0.37	0.17	0.16
		40	0.17	0.17	0.16	0.13	0.19	0.19
4	500	10	0.03	0.05	0.43	0.25	0.10	0.08
		20	0.03	0.05	0.12	0.06	0.09	0.07
		40	0.04	0.06	0.09	0.06	0.09	0.07
	1000	10	0.05	0.05	0.61	0.54	0.13	0.10
		20	0.05	0.06	0.17	0.10	0.10	0.09
		40	0.06	0.07	0.09	0.07	0.10	0.10
	2000	10	0.11	0.11	0.81	0.71	0.15	0.14
		20	0.11	0.12	0.35	0.25	0.16	0.15
		40	0.11	0.11	0.13	0.11	0.15	0.15
5	500	10	0.02	0.03	0.38	0.19	0.10	0.07
		20	0.04	0.06	0.12	0.06	0.08	0.06
		40	0.04	0.06	0.09	0.06	0.09	0.07
	1000	10	0.04	0.04	0.52	0.43	0.10	0.08
		20	0.05	0.07	0.16	0.09	0.10	0.09
		40	0.05	0.07	0.10	0.07	0.11	0.09
	2000	10	0.08	0.08	0.68	0.60	0.14	0.13
		20	0.09	0.10	0.28	0.18	0.13	0.13
		40	0.09	0.10	0.13	0.10	0.16	0.15

ítems politómicos exige el colapsamiento de celdillas con valores pequeños para evitar que se dispare el valor del índice. Este colapsamiento en tanto que se produce de forma diferente para las distintas filas, afecta a la estructura de la tabla de contingencia, lo cual posiblemente afecte a su vez a la distribución χ^2 supuesta. Adicionalmente, el algoritmo iterativo necesario para su cálculo resulta computacionalmente costoso, y más cuanto más largo es el test o más alternativas de respuesta presentan los ítems. Por último, al basarse en las puntuaciones totales de los sujetos, calculadas como la suma de sus aciertos en tests dicotómicos o como la suma de unos valores otorgados a las alternativas de respuesta en politómicos, el índice se comporta como si la puntuación del sujeto en el test fuera un estimador suficiente de su nivel en el rasgo latente, lo cual sólo es cierto en los modelos de Rasch. Esto implica también que el índice sólo puede utilizarse para evaluar el ajuste de ítems cuyo modelo permita que las categorías de respuesta puedan ordenarse por su contribución a la puntuación total.

Estudios posteriores deberían comprobar el funcionamiento de estos índices con un número de replicas superior. Aunque 100 es el número habitual utilizado en estudios con índices de ajuste en TRI, otras áreas nos muestran que un número mayor de réplicas (1000) permite la estimación de las tasas con una mayor precisión. Igualmente, aunque en este trabajo el desajuste se introdujo mediante modificaciones en los parámetros (siguiendo a Stone y Zhang, 2003), convendría comparar su funcionamiento en otras condiciones de desajuste, por ejemplo, cuando los datos han sido generados con un modelo y se pretende ajustarlos con otro (como hacen, por ejemplo, Orlando y Thissen, 2003).

Es necesario encontrar una alternativa para evaluar el ajuste de los modelos en TRI. Viendo cómo otras áreas (como la de los modelos de ecuaciones estructurales) han tratado de resolver el problema del ajuste, quizás haya que buscar la solución en medidas de tamaño del efecto. Índices derivados de los propuestos en otras áreas (que se basan la mayor parte de las veces a su vez en estadísticos χ^2, de los que como se ha podido ver disponemos en abundancia y variedad) o soluciones ofrecidas en regresión logística pueden ser las vías a explorar en los próximos años en el campo de los índices de ajuste en TRI.

Tabla 4
Tasa de detecciones correctas (potencia) de ítems desajustados para los distintos índices de ajuste (α= .05)

Número de alternativas, longitud del test (ítems) y tamaño de la muestra (N)			Métodos					
			Stone (2000)		Yen (1981), McKinley y Mills (1985)		Orlando y Thissen (2003)	
Alternativas	N	ítems	G^2*	X^2*	Q_1-G^2	Q_1-X^2	S-G^2	S-X^2
3	500	10	**0.96**	**0.94**	0.83	0.74	**0.43**	**0.31**
		20	**0.86**	**0.86**	0.69	**0.61**	**0.49**	**0.36**
		40	**0.74**	**0.69**	**0.60**	**0.48**	**0.51**	**0.40**
	1000	10	**1.00**	**1.00**	1.00	1.00	**0.83**	**0.77**
		20	**1.00**	**1.00**	0.94	0.93	0.87	0.89
		40	**1.00**	**1.00**	**0.85**	**0.85**	**0.82**	**0.83**
	2000	10	1.00	1.00	1.00	1.00	1.00	1.00
		20	1.00	1.00	1.00	1.00	1.00	1.00
		40	1.00	1.00	1.00	1.00	1.00	1.00
4	500	10	**0.96**	**0.95**	0.80	0.73	**0.54**	**0.42**
		20	**0.78**	**0.79**	0.65	**0.60**	**0.58**	**0.50**
		40	**0.69**	**0.67**	**0.62**	**0.53**	**0.58**	**0.48**
	1000	10	**1.00**	**1.00**	0.97	0.96	0.85	**0.86**
		20	**1.00**	**1.00**	0.90	0.88	**0.83**	0.84
		40	**0.98**	**0.98**	**0.84**	**0.84**	**0.80**	**0.81**
	2000	10	1.00	1.00	1.00	1.00	1.00	1.00
		20	1.00	1.00	1.00	1.00	1.00	1.00
		40	1.00	1.00	1.00	1.00	1.00	1.00
5	500	10	**0.83**	**0.84**	0.73	0.65	**0.56**	**0.47**
		20	**0.73**	**0.74**	0.68	**0.61**	**0.60**	**0.46**
		40	**0.66**	**0.63**	**0.64**	**0.56**	**0.61**	**0.50**
	1000	10	**1.00**	**1.00**	0.96	0.93	**0.85**	**0.88**
		20	**1.00**	**1.00**	0.88	**0.85**	**0.82**	**0.81**
		40	**0.92**	**0.92**	**0.81**	**0.80**	0.77	**0.78**
	2000	10	**1.00**	**1.00**	1.00	1.00	1.00	1.00
		20	**1.00**	**1.00**	1.00	1.00	1.00	1.00
		40	**1.00**	**1.00**	1.00	**1.00**	1.00	1.00

* En negrita: valores para los que la tasa de falsas alarmas es menor o igual que 0.10

Referencias

Ankenmann, R.D., Witt, E.A., y Dunbar, S.B. (1999) An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36*(4), 277-300.

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

DeMars, C.E. (2005). Type I error rates for PARSCALE's fit index. *Educational and Psychological Measurement*, *65*, 42-50.

Glas, C.A.W., y Suárez-Falcón, J.C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*(2), 87-106.

Kang, T., y Chen, T.T. (2007). *An investigation of the performance of the generalized S-X² item-fit index for polytomous IRT models*. ACT Research Report Series, 2007-1.

Lord, F.M., y Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score «equatings». *Applied Psychological Measurement, 8,* 452-461.

McKinley, R.L., y Mills, C.N. (1985). A comparasion of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*(1), 49-57.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 59-71.

Muraki, E., y Bock, R.D. (1997). *PARSCALE: IRT item analysis and test scoring for rating scale data [Computer software]*. Chicago: Scientific Software.

Olea, J., Ponsoda, V., y Prieto, G. (1999). *Tests informatizados: fundamentos y aplicaciones.* Madrid: Pirámide.

Orlando, M., y Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50-64.

Orlando, M., y Thissen, D. (2003). Further investigation of the performance of S - X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289-298.

Reise, S.P. (1990). A comparision of item and person-fit methods of assessing model data fit in IRT. *Applied Psychological Measurement, 14,* 127-137.

Roberts, J.S. (2008). Modified likelihood-based item fit statistics for the generalized graded unfolding model. *Applied Psychological Measurement, 32*, 407-423.

Samejima, R. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, 17.*

Stone, C.A., y Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40*(4), 331-352.

Stone, C.A. (2000) Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, *37*, 58-75.

Sueiro, M.J., y Abad, F.J. (en preparación). *MRFITIT: Goodness-of-fit software for IRT models.* Unpublished software.

Swaminathan, H., Hambleton, R.K., y Rogers, H.J. (2007). Assessing the fit of item response theory models, en C.R. Rao y S. Sinharay (Eds.): *Handbook of Statistics, vol. 26,* North Holland.

Thissen, D., Chen, W-H., y Bock, R.D. (2003). *Multilog (version 7) [Computer sotware]*. Lincolnwood, IL: Scientific Software International.

Thissen, D., Pommerich, M., Billeaud, K., y Williams, V.S. (1995) Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement. Special Issue: Polytomous item response theory, 19*(1), 39-49.

von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data. *Methods of Psychological Research Online, 2*(2)*,* 29-48.

Yen, W.M. (1981) Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*(2), 245-262.