Psicothema
2001. Vol. 13, nº 2, pp. 330-334

ISSN 0214 - 9915 CODEN PSOTEG

Un estudio de potencia y tasa de error tipo I del estadístico de razón de verosimilitud en la TRI

José Antonio López Pina y Mª Dolores Hidalgo Montesinos
Universidad de Murcia

En este estudio tratamos de probar la potencia (y tasa de error tipo I) del estadístico de razón de verosimilitud utilizado en BILOG (v. 3.04) para probar el ajuste de un modelo especificado a un ítem concreto. El modelo seleccionado es el modelo de 1-p. Para ello, utilizamos un tamaño de test fijo (75 ítems), y cuatro tamaños muestrales (100, 250, 500 y 1000). Además, utilizamos tres distribuciones de habilidad: una centrada N(0,1), y las otras dos no centradas: N(-1,1) y N(1,1). Para estudiar la potencia del estadístico de razón de verosimilitud hemos procedido a manipular el parámetro de discriminación en 25 ítems de los 75 que forman el test, de tal suerte que utilizamos seis niveles de discriminación media: (.3, .6, 1.0, 1.2, 1.5 y 1.8). Los resultados apuntan que la potencia del estadístico de razón de verosimilitud es muy baja cuando el tamaño muestral fue de sólo 100 sujetos, aun cuando el parámetro de discriminación fuera .3 ó 1.8. La potencia aumentó con el tamaño muestral apreciablemente, aunque el estadístico fue incapaz de detectar ítems cuyo parámetro de discriminación fue 1.0, 1.2 ó 1.5. Finalmente, la tasa de error tipo I se mantuvo dentro de los niveles nominales, excepto cuando el tamaño muestral fue elevado, tal como era de esperar.

A study of power and type I error rate to the likelihood ratio statistics in IRT. In this report we studied the power (and the type I error rate) of the likelihood ratio statistic used in BILOG (v. 3.04) to prove de fit of the model to the items. For this, we used a test size fixed (75 items) and four sample sizes (100, 250, 500 and 1000). Further, we used three ability distributions: one centered N(0,1) and another two noncentered: N(-1,1) and N(1,1). We also manipulated the mean discrimination in 25 of 75 items in six conditions (.3, .6, 1.0, 1.2, 1.5 and 1.8). The results showed that the power of the likelihood ratio statistic is low when the sample size was 100, although the discrimination parameter was .3 or 1.8. The power increased with the sample size, although this statistic cannot detect items that not follow the 1-p model, independently if the discrimination parameter was 1.0, 1.2 or 1.5. Finally, the type I error rate adjusted to the nominal levels, except when the sample size was high.

La aplicación de un modelo de respuesta al ítem lleva siempre aparejado un estudio de bondad de ajuste. Este estudio se aborda a través diversas perspectivas, desde la comprobación de los supuestos (unidimensionalidad e independencia local) hasta la realización de pruebas estadísticas para comprobar el ajuste del modelo a los ítems. Los beneficios que se pueden obtener con la aplicación de la Teoría de la Respuesta a los Items (TRI) sólo se podrán asegurar una vez que se ha realizado el estudio pertinente de bondad de ajuste; la validez del modelo depende de ello (Barbero, 1999; De Ayala, 1990; Hambleton, 1990).

Desde los inicios de la TRI, muchos investigadores han propuesto procedimientos para evaluar el ajuste del modelo ítem a ítem del test. En general, estos procedimientos estadísticos utilizan una distribución basada en χ^2. Así, Wright y Panchapakesan (1969), Wright y Mead (1977), y van den Wollenberg (1982a, b) fueron los primeros en proponer estadísticos de ajuste de los ítems para el modelo de Rasch. Bock (1972) propuso el estadístico χ^2_B para el modelo logístico de 2-p, y Yen (1981) propuso el estadístico Q_1 para los modelos logísticos de 1-p, 2-p y 3-p. Andersen (1973) propuso un estadístico de razón de verosimilitud en el contexto del modelo de Rasch, y Waller (1981) utiliza este mismo estadístico para evaluar el ajuste de cualquiera de los tres modelos de respuesta para ítems dicotómicos. Una revisión detallada de estos procedimientos aparece en Traub y Lam (1985; López-Pina e Hidalgo, 1996).

Sin embargo, algunos investigadores (Hambleton y Swaminathan, 1985; Hambleton, Swaminathan y Roger, 1991) desaconsejan la utilización de estos estadísticos a la hora de tomar decisiones sobre si un ítem sigue un modelo logístico o no. La causa fundamental reside en que cualquier estadístico basado en χ^2 presenta una potencia estadística baja con tamaños muestrales pequeños o una tasa de error tipo I muy elevada cuando el tamaño muestral es grande. Así, Hambleton et al. (1991; Hambleton y Murray, 1983) afirman que la tasa de error tipo I se incrementa con el tamaño muestral, lo que implica que cuando se utilizan grupos grandes o muy grandes (N=1000 o más), como por otra parte es necesario para la estimación de parámetros en los modelos de 2-p y 3-p, la probabilidad de

Correspondencia: José Antonio López Pina
Facultad de Psicología
Universidad de Murcia
30100 Murcia (Spain)
E-mail: jlpina@um.es

encontrar ítems que no se ajustan al modelo es mayor que la tasa nominal establecida (p.e. α=.05); la incidencia del incremento del tamaño muestral sobre la tasa de error tipo I puede ser tan elevada como para provocar que prácticamente todos los ítems sean rechazados como que no se ajustan al modelo especificado.

McKinley y Mills (1985) investigaron la tasa de error tipo I de cuatro estadísticos de bondad de ajuste en los modelos logísticos de 1-p, 2-p, 3-p, y un modelo analítico factorial. Los estadísticos estudiados fueron: χ^2 de Bock (1972), χ^2 de Yen (1981), χ^2 de Wright y Mead (1977) y un estadístico χ^2 de razón de verosimilitud de Bishop, Fienberg y Holland (1975). Para ello cruzaron tres distribuciones de habilidad (un grupo de baja habilidad, un grupo de alta habilidad y otro cuya habilidad estaba centrada con respecto a la dificultad media del test) con tres tamaños muestrales (500, 1000 y 2000) y los cuatro modelos, lo que resultó en 144 matrices simuladas que sirvieron como base para el análisis. McKinley y Mills (1985) concluyeron que el procedimiento de χ^2 basado en el estadístico de razón de verosimilitud produjo la tasa de error tipo I más baja, mientras que el estadístico de Bock (1972) fue el que tuvo la menor tasa de error tipo II. Además, el hecho de utilizar una grupo de habilidad baja, media o alta provocó ligeras fluctuaciones, aunque en su opinión carecieron de importancia. En cualquier caso, las tasas de error tipo I fueron mayores cuando la habilidad del grupo fue media que cuando fue alta o baja, independientemente del tamaño muestral y del modelo estudiado.

La temprana detección de los problemas asociados a la toma de decisiones cuando se utilizan los estadísticos de bondad de ajuste de los ítems ha prevenido sobre su uso, aunque los programas informáticos de TRI siguen manteniéndolos en las salidas de estimación de parámetros para que sirvan como un primer índice de detección de un posible mal ajuste del ítem. Por ello, a pesar de que el estadístico de razón de verosimilitud parece que se puede utilizar razonablemente en la toma de decisiones, sería conveniente realizar un estudio de alguna de las implementaciones más modernas de este estadístico. Para realizar este estudio se seleccionó el estadístico de razón de verosimilitud basado en χ^2 implementado en BILOG (Mislevy y Bock, 1990), dado que es uno de los programas informáticos más extendidos tanto en el campo aplicado como de investigación en TRI.

Para evaluar la potencia y tasa de error tipo I de este estadístico decidimos manipular el supuesto de CCIs paralelas del modelo logístico de 1-p, ya que este modelo es más restrictivo en sus supuestos, y por tanto más fácil detectar los ítems que no se ajustan al modelo.

El estadístico de razón de verosimilitud

En este estudio se ha utilizado el estadístico de razón de verosimilitud que aparece en el programa BILOG (v. 3.04) (Mislevy y Bock, 1990). Esta versión implementa el método de máxima verosimilitud marginal para estimar los parámetros de habilidad y de los ítems en los modelos logísticos de 1-p, 2-p y 3-p. La superioridad de este método sobre otros, como el método de máxima verosimilitud conjunta ha sido puesta de manifiesto en diversas investigaciones (Seong, 1990; Stone, 1992), ya que el uso de distribuciones previas mejora sustancialmente la estimación de parámetros de los ítems y de la habilidad.

El programa BILOG dispone de tres procedimientos para evaluar la bondad de ajuste de los ítems, dependiendo del tamaño del test. Así, si el test tiene menos de 10 ítems, BILOG emplea un estadístico de razón de verosimilitud, basado en χ^2, que prueba el ajuste del modelo al conjunto total del test antes que a los ítems individualmente. Si el test tiene entre 11 y 20 ítems, BILOG calcula un residual a posteriori estandarizado sobre los puntos cuadratura definidos en el proceso de estimación de parámetros. Desafortunadamente, no existe una distribución conocida para este estadístico que permita una prueba estadística de significación. Por último, si el test tiene más de 20 ítems BILOG propone utilizar un estadístico de razón de verosimilitud, también basado en χ^2, que se obtiene de comparar las frecuencias de respuestas correctas e incorrectas observadas dentro de cada uno de los k intervalos seleccionados con las frecuencias de respuestas correctas e incorrectas esperadas para la media del intervalo (μ_{θ_k}), según el modelo que se desee ajustar. Este estadístico se expresa como:

$$G_j^2 = 2\sum_{k=1}^{n_k}\left[r_{jk}\log_e \frac{r_{jk}}{N_k P_j\left(\mu_{\theta_k}\right)} + \left(N_k - r_{jk}\right)\log_e \frac{N_k - r_{jk}}{N_k\left[1 - P_j\left(\mu_{\theta_k}\right)\right]}\right]$$

donde n_k es el número de intervalos, r_{kj} es la proporción de respuestas correctas en el k intervalo al ítem j. N_k es el número de sujetos en el intervalo y $P_j(\mu_{\theta_k})$ es la probabilidad de acertar el ítem j, según el modelo de respuesta al ítem seleccionado, del sujeto cuya habilidad sea igual a la habilidad media del intervalo. Los grados de libertad son equivalentes al número de intervalos seleccionados, ya que las estimaciones de máxima verosimilitud empleadas para estimar $\hat{\theta}$ minimizan el valor del χ^2 calculado.

Método

Condiciones experimentales

La potencia de los estadísticos basados en χ^2 depende del tamaño muestral, por lo que hemos seleccionado cuatro tamaños muestrales (100, 250, 500 y 1000) para observar la incidencia del mismo en la potencia del estadístico de razón de verosimilitud. Para cada uno de los cuatro tamaños muestrales se simularon tres distribuciones normales donde una representó a un grupo de habilidad media N(0,1), otra a un grupo de baja habilidad N(-1,1), y la última a un grupo de alta habilidad N(1,1), con la intención de probar si la potencia del estadístico de razón de verosimilitud se ve afectada en función de la habilidad del grupo donde se calibra el test. La longitud del test se fijó en 75 ítems en todas las condiciones experimentales, ya que consideramos que esta dimensión no es relevante en el estudio que nos ocupa, aunque este número de ítems nos permite obtener suficientes réplicas para evaluar tanto la potencia como la tasa de error tipo I del estadístico de razón de verosimilitud implementado en BILOG. Los parámetros de dificultad de los 75 ítems fueron los mismos en todas las condiciones, y se generaron bajo una distribución uniforme en el intervalo (-2, 2).

Dado que el estudio de potencia se realiza sobre el modelo de 1-p, manipulamos los parámetros de discriminación en 25 ítems del test. Para ello, empleamos seis niveles de discriminación media (.3, .6, 1.0, 1.2, 1.5 y 1.8), donde en cinco de ellas (.3, .6, 1.2, 1.5 y 1.8) se viola claramente el supuesto de iguales parámetros de discriminación para todos los ítems del test. Las medias y desviaciones típicas de los parámetros de discriminación en las seis condiciones aparecen en la tabla 1. En los 50 ítems restantes mantuvimos el parámetro de discriminación fijado (a=1).

Tabla 1
Condiciones experimentales del parámetro de discriminación

Condición	μ_a	σ_a
1	.292	.065
2	.595	.070
3	.997	.056
4	1.206	.060
5	1.503	.058
6	1.792	.065

Los cuatro tamaños muestrales, las tres distribuciones de habilidad y los seis niveles de discriminación suponen un total de 72 condiciones experimentales. En cada una de estas condiciones se realizaron 40 réplicas, por lo que disponemos de 1000 estadísticos por condición experimental para probar la potencia del estadístico de razón de verosimilitud.

Para determinar la tasa de error tipo I se generaron nuevas matrices de respuestas utilizando los 75 ítems originales, pero con el parámetro de discriminación fijado en 1. Cada condición experimental resultante de cruzar los 4 tamaños muestrales con las 3 distribuciones de habilidad se replicó en 40 ocasiones, con lo que disponemos de 36000 estadísticos (3000 en cada condición) para obtener esta tasa.

Procedimiento

Cada una de las matrices resultantes de las condiciones experimentales en este estudio fueron simuladas con un programa realizado para tal fin siguiendo las pautas de Hambleton y Cook (1983). Este programa calcula, en el modelo logístico de 2-p, la probabilidad de que un sujeto de habilidad θ acierte el ítem. Entonces, compara esta probabilidad con un número generado aleatoriamente. Si la probabilidad es menor que el número aleatorio generado, se considera que la respuesta del sujeto es un acierto (1). En caso contrario, es un fallo (0). Las réplicas dentro de cada condición fueron generadas con los mismos parámetros, aleatorizando la semilla del generador de números aleatorios del programa informático.

Una vez construidas las matrices de respuestas a los ítems se realizaron las estimaciones de los parámetros de los ítems y de la habilidad con BILOG (v. 3.04), utilizando las opciones por defecto de los comandos de este programa. No obstante, nuestro objetivo no fueron los parámetros estimados en sí mismos, sino los estadísticos de razón de verosimilitud que permiten comprobar el ajuste del modelo al ítem.

Resultados

Potencia del estadístico de razón de verosimilitud

La tabla 2 presenta la potencia del estadístico de razón de verosimilitud en función del tamaño muestral y el nivel de significación en los seis niveles de discriminación seleccionados para el grupo de habilidad media N(0,1). Se observa, al nivel de significación del 5%, que la potencia creció conforme aumentó el tamaño muestral y se alejó la discriminación media de los ítems que no se correspondían con el modelo especificado. En este sentido conviene destacar que cuando el tamaño muestral fue de 100 sujetos, el estadístico de razón de verosimilitud presentó una potencia muy baja, independientemente de la discriminación promedio de los 25 ítems, destacando que en el intervalo de discriminación central (.6 a 1.5), el estadístico fue incapaz de detectar los ítems que se habían generado bajo el modelo logístico de 2-p.

En esta tabla también se aprecia (n.s.: 5%) como el incremento del tamaño muestral produce un incremento paralelo en la potencia del estadístico de razón de verosimilitud, apreciable sólo en las condiciones más extremas. Así, para μ_a=.3, la potencia del estadístico fue .874 (N=1000), para μ_a=1.5 la potencia fue .734 (N=1000), y para μ_a=1.8 fue .933, en el mismo tamaño muestral.

Sin embargo, cuando los parámetros de discriminación promedio fueron de μ_a=.6, μ_a=1.0 y μ_a=1.2, la potencia del estadístico estuvo muy limitada (.303, .069 y .230, respectivamente), lo que revela que aun con tamaños muestrales elevados, si el parámetro de discriminación del ítem está muy cercano al valor supuesto por el modelo de 1-p (a=1), el estadístico de razón de verosimilitud no es un buen estadístico para probar el ajuste del ítem bajo el modelo de 1-p.

En la misma tabla aparece la potencia del estadístico en los niveles de significación α=.01 y α=.001. Como era esperable, la potencia bajó sustancialmente, destacando que la utilización de niveles de significación muy extremos (α=.001), cuando el tamaño muestral es igual o menor a 500 sujetos, puede llevarnos a incluir ítems en un test que sigue el modelo de 1-p, aun cuando realmente su capacidad discriminativa difiera sustancialmente de lo establecido por el modelo.

La tabla 3 presenta la potencia del estadístico de razón de verosimilitud cuando la habilidad media del grupo fue baja N(-1,1). Como se aprecia en la misma, los resultados son equivalentes al caso donde la habilidad del grupo fue media, aunque la potencia fue algo menor en todas las condiciones. Sin embargo, la pérdida de potencia de la prueba fue más apreciable cuando el parámetro de discriminación fue elevado (μ_a=1.5) o muy elevado (μ_a=1.8) y el tamaño muestral de 500 o 1000 sujetos, que cuando el parámetro de discriminación fue bajo (μ_a=.3), en los mismos tamaños muestrales. En el primer caso, para N=500 y μ_a=1.5, la diferencia entre ambas condiciones en potencia fue de .165, mientras que cuando μ_a=1.8, esta diferencia es de .225. Sin embargo, cuando μ_a=.3, la diferencia en potencia fue de .030. Para N=1000, las diferencias en μ_a=1.8 y μ_a=1.5, fueron .148 y .191, mientras que para μ_a=.3, la diferencia en potencia fue de .074.

Tabla 2
Potencia del estadístico de razón de verosimilitud
Distribución centrada N(0,1)

Nivel de significación	Tamaño muestral	Parámetro de discriminación					
		μ_a=.3	μ_a=.6	μ_a=1	μ_a=1.2	μ_a=1.5	μ_a=1.8
α = .05	100	.179	.073	.039	.038	.054	.118
	250	.384	.121	.046	.062	.170	.358
	500	.610	.166	.064	.121	.406	.724
	1000	.874	.303	.069	.230	.734	.933
α = .01	100	.076	.019	.008	.005	.013	.023
	250	.181	.041	.006	.010	.056	.144
	500	.405	.053	.008	.037	.187	.489
	1000	.756	.147	.014	.094	.514	.864
α = .001	100	.015	.002	.000	.000	.000	.001
	250	.052	.004	.000	.001	.008	.040
	500	.192	.010	.002	.008	.061	.249
	1000	.579	.055	.001	.026	.267	.707

La tabla 4 presenta la potencia del estadístico de razón de verosimilitud para el grupo de alta habilidad N(1,1). En general, los resultados son similares a los obtenidos con los dos grupos anteriores, altas tasas de potencia con tamaño muestral elevado (N=1000, μ_a=1.8, 1-ß =.865) y bajas, independientemente del tamaño muestral cuando la discriminación promedio estuvo cercana al valor supuesto del parámetro de discriminación por el modelo (N=1000, μ_θ, 1-ß=.058).

En términos generales, en los tamaños muestrales de 100 y 250 sujetos, los resultados indican que en el grupo de habilidad media N(0,1), la potencia del estadístico fue mayor que en los grupos de habilidad baja N(-1,1) o alta N(1,1), aunque dada la baja potencia del estadístico, no se puede establecer una conclusión definitiva.

Tasa de error tipo I

La tabla 5 presenta las tasas de error tipo I bajo las tres distribuciones de habilidad. En los tres casos se aprecia que se controló relativamente bien la tasa de error tipo I hasta el tamaño muestral de 500 sujetos, incluso por debajo del nivel nominal cuando el tamaño muestral fue tan pequeño como 100 sujetos (p=.033, p=.029 y p=.033, respectivamente en las tres distribuciones de habilidad). Sin embargo, cuando N=1000, la tasa de error tipo I se aleja cierta cantidad del nivel nominal (α =.05), indicando que con tamaños muestrales elevados tenemos cierto riesgo de rechazar ítems como no ajustados al modelo especificado, aun cuando realmente sigan el modelo de 1-p.

Tabla 3
Potencia del estadístico de razón de verosimilitud
Distribución centrada N(-1,1)

Nivel de significación	Tamaño muestral	Parámetro de discriminación					
		μ_a=.3	μ_a=.6	μ_a=1	μ_a=1.2	μ_a=1.5	μ_a=1.8
	100	.150	.061	.040	.023	.042	.064
α = .05	250	.281	.115	.040	.050	.119	.218
	500	.580	.159	.051	.109	.241	.499
	1000	.800	.274	.070	.197	.543	.785
	100	.054	.018	.005	.003	.006	.016
α = .01	250	.125	.030	.006	.007	.024	.077
	500	.373	.064	.018	.027	.107	.294
	1000	.648	.116	.013	.076	.376	.666
	100	.011	.000	.000	.000	.001	.000
α = .001	250	.039	.003	.001	.001	.007	.019
	500	.159	.019	.004	.007	.040	.113
	1000	.457	.042	.002	.019	.183	.524

Tabla 4
Potencia del estadístico de razón de verosimilitud
Distribución centrada N(1,1)

Nivel de significación	Tamaño muestral	Parámetro de discriminación					
		μ_a=.3	μ_a=.6	μ_a=1	μ_a=1.2	μ_a=1.5	μ_a=1.8
	100	.166	.070	.043	.034	.061	.075
α = .05	250	.289	.092	.045	.076	.164	.301
	500	.599	.168	.042	.139	.384	.627
	1000	.846	.287	.058	.234	.648	.865
	100	.044	.014	.008	.009	.007	.012
α = .01	250	.153	.019	.011	.015	.055	.132
	500	.376	.064	.010	.044	.174	.407
	1000	.701	.132	.015	.098	.440	.763
	100	.007	.000	.002	.000	.000	.001
α = .001	250	.054	.001	.002	.001	.007	.030
	500	.174	.014	.001	.012	.040	.195
	1000	.480	.042	.001	.021	.210	.610

Tabla 5
Tasa de error tipo I
Distribución centrada N(1,1)

Nivel de Significación	Tamaño Muestral	Nivel de significación		
		α=.05	α=.01	α=.001
N(0,1)	100	.033	.004	.000
	250	.040	.007	.000
	500	.045	.011	.002
	1000	.059	.011	.002
N(-1,1)	100	.029	.004	.000
	250	.045	.009	.001
	500	.049	.013	.002
	1000	.062	.013	.002
N(1,1)	100	.033	.004	.000
	250	.039	.009	.000
	500	.055	.010	.001
	1000	.065	.016	.000

Discusión

Los resultados de un estudio de simulación son siempre limitados pero permiten informar sobre aspectos de los estadísticos que no siempre son aparentes en los estudios con datos reales. En este caso, es visible el hecho de que el estadístico de razón de verosimilitud tiene una potencia razonable siempre que el tamaño muestral sea mayor de 500 sujetos y el parámetro de discriminación se encuentre muy alejado del valor establecido en el modelo de 1-p. Así, para que el estadístico de razón de verosimilitud detecte que un ítem no se ajusta al modelo de 1-p, su parámetro de discriminación debe ser muy extremo (.33 o 1.8), sobre todo cuando el tamaño muestral empleado para calibrar el ítem sea tan bajo como 250 y/o 100 sujetos. En este caso, la potencia del estadístico es tan baja que el ítem será incluido en el modelo aun cuando su pendiente difiera sustancialmente de los otros ítems que sí siguen el modelo de 1-p.

Por regla general, se desaconseja la utilización de los estadísticos de ajuste de ítems basados en χ^2 para tomar la decisión sobre si un ítem debe o no incluirse en un test bajo un modelo especificado, máxime si el tamaño muestral es tan elevado como 1000 sujetos. Los resultados de este estudio reflejan, sin embargo, que en las condiciones experimentales empleadas, cuando N=1000, el estadístico de razón de verosimilitud controla razonablemente bien la tasa de error tipo I (p=.059, p=.062, y p=.065, en las tres distribuciones empleadas), y muestra la potencia más elevada para detectar el ajuste del modelo, por lo que existe un equilibrio razonable entre ambos aspectos que no abonan la extendida idea de que tamaños muestrales elevados conducen inevitablemente a tasas de error tipo I altas, al menos con este estadístico.

Los resultados encontrados en este estudio son similares a los obtenidos en el trabajo de McKinley y Mills (1985). Así, en las

condiciones donde los parámetros de discriminación son extremos (μ_a=0.3 y μ_a=1.8) y las de mayor tamaño muestral (500 y 1000 sujetos), condiciones comparables a las del trabajo de McKinley y Mills (1985), observamos que las tasas de potencia de los diferentes estadísticos de ajuste y las del estadístico de razón de verosimilitud son similares en ambos trabajos. Así, cuando la distribución de habilidad se encontró por debajo de la media N(1,1), McKinley y Mills (1985) obtuvieron unas tasas de potencia (n.s.=0.01) entre 0.28 y 0.36 (N=500), entre 0.45 y 0.47 (N=1000), y entre 0.61 y 0.64 (N=2000), mientras que en nuestro estudio estas tasas fueron, 0.37 (N=500 y μ_a=0.3), 0.29 (N=500 y μ_a=1.8), 0.65 (N=1000 y μ_a=0.3) y 0.67 (N=1000 y μ_a=1.8). Cuando la habilidad media del grupo N(0,1) fue igual a la dificultad media del test, McKinley y Mills (1985) encontraron unas tasas de potencia (n.s.=0.01) entre 0.36 y 0.40, para un tamaño muestral de 500 sujetos, y entre 0.49 y 0.53 para un tamaño muestral de 1000 sujetos, siendo las encontradas en el presente trabajo 0.41 (μ_a=0.3) y 0.49 (μ_a=1.8) cuando N=500, y 0.76 (μ_a=0.3) y 0.86 (μ_a=1,8) cuando N=1000. Por último cuando la media de habilidad del grupo N(1,1) se encontró por encima de la media de dificultad, las tasas de potencia estuvieron entre 0.40 y 0.47 (N=500), 0.63 y 0.69 (N=1000) y 0.75 y 0.77 (N=2000) en el estudio de McKinley y Mills (1985), y fueron de 0.38 (N=500 y μ_a=0.3), 0.41 (N=500 y μ_a=1.8), 0.70 (N=1000 y μ_a= 0.3) y 0.76 (N=1000 y μ_a=1.8), siendo la potencia del estadístico de razón de verosimilitud, como en el estudio de McKinley y Mills, ligeramente mayor en el grupo de habilidad por encima de la media que en los dos grupos restantes. Estos resultados sugieren la idea de que la potencia del estadístico es mayor cuando la habilidad media del grupo supera la dificultad media del test, ya que un mayor porcentaje de sujetos se concentra en los niveles más elevados de habilidad, proporcionando más información para detectar los ítems más discriminativos

El modelo de 1-p es muy restrictivo en las condiciones impuestas a los ítems, por lo que una idea muy extendida es que el ajuste del modelo a los datos de un test puede producir una elevada tasa de ítems no aceptados, que deben eliminarse o como mínimo ser reformulados. Este estudio, bajo las condiciones experimentales empleadas, muestra que el rechazo de ítems que no siguen el modelo, con el estadístico de razón de verosimilitud, no es tan fácil, sobre todo cuando el tamaño muestral es relativamente bajo (N<250), donde, por otra parte, se mueven gran parte de los estudios empíricos realizados con este modelo, por lo que debería extremarse el cuidado a la hora de seleccionar ítems con el estadístico de razón de verosimilitud si el tamaño muestral es bajo, situación que no se tiene en cuenta al considerar que el modelo de 1-p puede obtener estimaciones razonables de los parámetros de los ítems y de la habilidad de los sujetos con pequeños tamaños muestrales.

En definitiva, de este estudio se deduce que si en el proceso de selección de ítems se le quiere dar cierta importancia al estadístico individual de ajuste del ítem, parece aconsejable utilizar tamaños muestrales de 500 o más sujetos, aun cuando el modelo que se pretende ajustar sea el modelo de 1-p, ya que es a partir de este tamaño muestral cuando se obtiene el mejor equilibrio entre potencia y tasa de error tipo I.

Referencias

Andersen, E.B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123-140.

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

Bishop, Y.M.M., Fienberg, S.E. y Holland, P.W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: The MIT Press.

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

De Ayala, R.J. (1990). Assessing the fit of dichotomous ítem response theory models. *Behavior Research Methods, Instruments, & Computers, 22*, 80-81.

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

Hambleton, R.K. y Cook, L.L. (1983). Robustness of ítem 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: Latent trait test theory and computerized adaptive testing*, p. 31-49. New York: Vancouver.

Hambleton, R.K. y Murray, L. (1983). Some goodness of fit investigations for ítem response models. En R.K. Hambleton (Ed.). *Applications of ítem response theory* (pp. 71-94). Vancouver: Education Research Institute of British Columbia.

Hambleton, R.K. y Swaminathan, H. (1985). *ítem response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hambleton, R.K., Swaminathan, H. y Rogers, H.J. (1991). *Fundamentals of ítem response theory*. Newbury Park, CA: Sage Publications.

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

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. y Bock, R.D. (1990). *PC-BILOG 3.04: ítem analysis and test scoring with binary logistic models*. Moresville, IN: Scientific Software.

Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of ítem and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*, 299-312.

Stone, C. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*, 1-16.

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

van den Wollenberg, A.L. (1982a). A simple and effective method to test the dimensionality axiom of the Rasch model. *Applied Psychological Measurement, 6*, 83-91.

van den Wollenberg, A.L. (1982b). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-139.

Waller, M.I. (1981). A procedure for comparing logistic latent trait models. *Journal of Educational Measurement, 18*, 119-125.

Wright, B.D. y Mead, R.J. (1977). *BICAL: Calibrating items and scales with the Rasch model* (Research Memorandum No. 23). Chicago IL: University of Chicago, Statistical Laboratory, Department of Education.

Wright, B.D. y Panchapakesan, N.A. (1969). A procedure for sample-free ítem analysis. *Educational and Psychological Measurement, 29*, 23-37.

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

Aceptado el 7 de noviembre de 2000