Psicothema 2004. Vol. 16, nº 2, pp. 317-324
www.psicothema.com
ISSN 0214 - 9915 CODEN PSOTEG

SOFTWARE, INSTRUMENTACIÓN Y METODOLOGÍA

Influencia del sesgo de la distribución de habilidad en la distribución del estadístico l_z

Rosa Mª Núñez Núñez y José A. López Pina*
Universidad Miguel Hernández y * Universidad de Murcia

El estadístico de medición apropiada l_z de Drasgow, Levine y Williams (1985) es un índice adecuado para detectar patrones atípicos de respuesta por su alta tasa de identificaciones correctas. Sin embargo, algunos estudios han comprobado que la distribución normal de este índice está afectada, entre otros factores, por la longitud del test, el modelo de respuesta a los ítems o por la distribución de habilidad. En este trabajo se ha estudiado el efecto que tiene la simetría de la distribución de habilidad al tiempo que se manipulan la longitud del test, la magnitud del parámetro de discriminación y el procedimiento de estimación de los parámetros de habilidad y de los ítems. Los resultados constatan que el índice l_z sigue una distribución aproximadamente normal aunque sesgada y moderadamente leptocúrtica; las tasas de falsos positivos atestiguan que es una prueba conservadora y consistente en el nivel nominal de .05.

Influence of the ability distribution skewness on the distribution of statistic l_z. The appropriateness measurement statistic l_z of Drasgow, Levine & Williams (1985) is a suitable index for detecting aberrant patterns because of its high hit rates. However, the normal distribution of this index is affected, for instance, by the test length, the item response model or the ability distribution. This research analyses the effect of the ability distribution skewness on the distribution of l_z, and the test length, the extent of the discrimination parameter and the estimation process of the ability and items are also manipulated. The results show that the distribution of the index l_z is aproximately a normal distribution but skewed and sightly leptokurtic; the false positive rates point out that the index l_z is a conservative and consistent test in the significance level of .05.

La presencia de patrones atípicos de respuesta (PAR) repercute negativamente en la construcción de tests y de bancos de ítems con propiedades psicométricas, así como en el análisis de validez de los mismos (Drasgow y Guertler, 1987; Meijer, 1997, 1998; Schmitt, Chan, Sacco, McFarland y Jennings, 1999; Schmitt, Cortina y Whitney, 1993). La parte de la Psicometría encargada de identificar y tratar los patrones de respuesta atípicos fue definida por Levine y Rubin (1979) como *medición apropiada* dentro de la Teoría de Respuesta a los Ítems (TRI); recientemente Meijer y Sijtsma (1999, 2001) han denominado *métodos de ajuste de persona* a los métodos y estadísticos que identifican PAR debido o bien a la ausencia de ajuste de éstos con relación a un modelo de TRI o bien a la falta de concordancia entre dicho patrón y los patrones de la muestra a la que pertenece el sujeto. De todos ellos, el estadístico basado en la función de verosimilitud l_z de Drasgow, Levine y Williams (1985) ha sido muy utilizado dada su alta tasa de identificaciones correctas.

Fecha recepción: 22-1-03 • Fecha aceptación: 14-10-03
Correspondencia: Rosa Mª Núñez Núñez
Facultad de CC. Jurídicas y Sociales
Universidad Miguel Hernández
03202 Elche (Spain)
E-mail: rnunez@umh.es

El estadístico l_z

El estadístico l_z es la expresión estandarizada de l_0, índice propuesto por Levine y Rubin (1979) para detectar PAR, basado en la función de verosimilitud:

$$l_0 = V\left(u|\theta,a,b,c\right) = \sum_{i=1}^{N}\sum_{j=1}^{n}\left[u_{ij}\,1nP_j\left(\theta_i\right) + \left(1+u_{ij}\right)1nQ_j\left(\theta_i\right)\right]$$

donde u es el patrón de respuestas de un sujeto de habilidad θ en un test de n ítems dicotómicos; u_{ij} es la respuesta del sujeto i al ítem j, la cual es 1 si acierta el ítem y 0 si lo falla; $P_j(\theta_i)$ es la pro-

babilidad de acertar el ítem *j* por el sujeto *i*, y $Q_j(\theta_i)$ es la probabilidad de fallarlo. Sin embargo, aunque comprobada la eficacia de l_0 para detectar PAR (Levine y Drasgow, 1982; Levine y Rubin, 1979), este índice presentaba dos problemas relevantes: el primero, no está estandarizado con lo cual clasificar un patrón de atípico o no depende de θ; el segundo, se desconoce la distribución de l_0, característica esencial para poder probar la hipótesis nula de que un patrón de respuestas es normal. Drasgow, Levine y Williams (1985) resolvieron estos dos inconvenientes estandarizando l_0 de acuerdo con la siguiente expresión:

$$l_z = \frac{l_0 - E(l_0)}{\sigma(l_0)}$$

donde $E(l_0)$ es el valor esperado de l_0:

$$E(l_0) = \sum_{j=1}^{n}\left[P_j(\theta)1nP_j(\theta) + Q_j(\theta)1nQ_j(\theta)\right]$$

y $\sigma(l_0)$ es la desviación típica de l_0:

$$\sigma(l_0) = \sqrt{\sum_{j=1}^{n} P_j(\theta)Q_j(\theta)\left[1n\frac{P_j(\theta)}{Q_j(\theta)}\right]^2}$$

Con l_z se obtiene el máximo de la función de verosimilitud una vez que han sido estimados los parámetros de los ítems y de la habilidad que se ajustan al modelo de respuesta a los ítems (MRI). Este estadístico sigue una distribución normal de media 0 y desviación típica 1 cuando los datos se ajustan al modelo. Si los valores de l_z se aproximan a 0, el patrón de respuesta es apropiado; si l_z tiene valores negativos, el patrón es atípico; y si el estadístico l_z es positivo, el patrón de respuesta observado es más apropiado que el pronosticado por el modelo.

Distintas investigaciones han comprobado que la normalidad del estadístico está afectada por el método de estimación de la habilidad (Meijer y Nering, 1997; Molenaar y Hoijtink, 1990; Nering, 1995; Reise, 1995), la longitud del test (Nering, 1995; Noonan, Boss y Gessaroli; 1992; Reise y Flannery, 1996), el tipo de test (Nering, 1997; Reise, 1995), el modelo de respuesta ajustado (Noonan *et al.*, 1992), la dimensionalidad del test (Li y Olejnik, 1997) y por la presencia de patrones atípicos (Drasgow, Levine y McLaughlin, 1987; Noonan *et al.*, 1992; Reise, 1995). En este trabajo de simulación se ha estudiado el efecto de la simetría de la distribución de habilidad junto con la manipulación del número de ítems del test, la magnitud del parámetro de discriminación y el procedimiento de estimación de los parámetros de habilidad y de los ítems, cuando se emplea el modelo logístico de 2-p para ajustar los datos de un test o banco de ítems.

Método

Se ha generado una matriz de 1.000 patrones de respuesta por el algoritmo de Hambleton y Cook (1983) según el cual, una vez elegido el MRI (2-p), requiere las siguientes especificaciones: el número de sujetos de la muestra (N), la distribución de habilidad, el número de ítems (n) y los valores de los parámetros de habilidad y de los ítems; las longitudes de test son 10, 25, 50 y 75 ítems. Con objeto de valorar si la simetría de la distribución de θ afecta a la distribución de l_z, se ha trabajado con tres distribuciones de θ normales N(0,1): no sesgada, asimétrica positiva con índice de sesgo +1 y asimétrica negativa con índice de sesgo -1. Los parámetros de los ítems proceden del trabajo de Narayanan y Swaminathan (1996) en el que el test original estaba formado por 40 ítems dicotómicos ajustados al modelo de 3-p. En este estudio el parámetro de pseudo-azar se mantuvo constante e igual a 0. La posible influencia del parámetro de discriminación se valoró incrementando los parámetros originales en magnitudes de .30 y .60, ocasionando así tres condiciones experimentales: (C1) valores de a_j verdaderos; (C2) valores de a_j verdaderos con incremento de .30; y (C3) valores de a_j verdaderos con incremento de .60.

Para analizar el efecto del método de estimación de los parámetros de la habilidad y de los ítems sobre la distribución de l_z se han empleado dos procedimientos: estimación por máxima verosimilitud marginal (MV) y estimación esperada a posteriori (EAP), llevadas a cabo con el programa BILOG v. 3.04 (Mislevy y Bock, 1990). El análisis del método de estimación se ha acompañado de un estudio de recubrimiento de los parámetros θ mediante el coeficiente de correlación de Pearson (ρ), la exponencial de la raíz del error cuadrático medio (RMSE) y el error medio con signo (ASB).

La distribución de l_z se ha descrito con los estadísticos media, desviación típica, sesgo y curtosis. Se han completado estos resultados con la prueba de normalidad de Lilliefors (1967; Marascuilo y McSweeney, 1977), una variante de la prueba de Kolmogorov-Smirnov útil cuando la media y la desviación típica de la distribución son desconocidas. Esta prueba no está afectada ni por la localización ni por la escala de l_z y el contraste lo realiza según la forma de la distribución por aproximación no lineal a la tabla de Lilliefors. Tanto los estadísticos descriptivos como la prueba no paramétrica de Lilliefors se han calculado con el programa SYSTAT v. 10.0 (2000).

También se han evaluado las tasas de falsos positivos (FP) del índice l_z para un contraste bilateral en dos niveles de significación: .05 y .01; en el caso de que la distribución de dicho estadístico no sea la normal tipificada, las tasas de FP bien serán infraestimadas o bien sobrestimadas.

Resultados

Estudio de recubrimiento

Si bien es cierto que el estadístico l_z está estandarizado y, por lo tanto, es independiente de los niveles de θ en los que se calcule, la distribución de θ y el empleo de valores de habilidad verdaderos o estimados no deberían repercutir en la distribución del estadístico. Analizando $\rho(\theta,\hat{\theta})$ y el índice RMSE, las estimaciones de θ en las condiciones experimentales que contenían ítems más discriminativos mejoraban tanto con MV como con EAP; el poder discriminativo de los ítems tuvo mayor influencia en los tests más cortos (10 y 25 ítems), donde $\rho(\theta,\hat{\theta})$ y RMSE se incrementaron notablemente desde la condición de menor discriminación (C1) a la de mayor discriminación (C3). Otro factor relacionado con la mejoría en la estimación de θ es la longitud del test, por la que, con el aumento en número de ítems se ganó en grado de acuerdo con los parámetros de sujeto.

El índice ASB muestra que ambos procedimientos de estimación tienden a sobrestimar la habilidad real sobre todo en el test más largo (75 ítems) y estimando con MV, así como cuando la distribución de habilidad no está sesgada. El parámetro a_j no tuvo un efecto definido.

Lo cierto es que las diferencias entre MV y EAP no son muy acusadas y, por lo tanto, la elección de un procedimiento u otro no

debería contaminar el cálculo de l_z. Estas mismas conclusiones referentes al recubrimiento del parámetro de habilidad coinciden con las que obtuvieron Meijer y Nering (1997), y Nering (1995). El aumento del número de ítems y el incremento del parámetro a_j sí optimizan las estimaciones, lo que también concuerda con el estudio de Hambleton y Cook (1983).

A continuación se presentan los estadísticos descriptivos de l_z antes citados, la prueba de Lilliefors y de las tasas de FP; debido a que son muchos los datos que se obtienen y con objeto de simplificar y facilitar la interpretación de los mismos al lector, se ha optado por exponer los resultados más representativos.

Tabla 1
Estadísticos descriptivos de l_z, prueba de normalidad y FP con distribución de habilidad no sesgada

θ	n	C	Media	DT	Sesgo	Curtosis	M.D.	p	FP .05	FP .01	N[a]
Parámetros verdaderos	10	1	.037	1.002	-.858**	.482**	.099**	.000	.039	.013	1000
		2	.013	.992	-1.243**	2.751**	.081**	.000	.044	.017	1000
		3	-.014	1.048	-1.281**	2.018**	.104**	.000	.041	.022	1000
	25	1	-.057	1.033	-.337**	-.171	.037**	.002	.039	.006	1000
		2	-.022	.998	-.402**	-.106	.039**	.001	.041	.008	1000
		3	-.038	1.026	-.654**	.506**	.064**	.000	.037	.011	1000
	50	1	-.040	1.026	-.311**	-.071	.035**	.006	.041	.009	1000
		2	.010	1.019	-.318**	.084	.035**	.006	.055	.007	1000
		3	-.063	1.061	-.498**	.324*	.055**	.000	.046	.015	1000
	75	1	.070	.993	-.422**	.678**	.041**	.001	.056	.010	1000
		2	.060	.985	-.318**	.147	.033*	.014	.054	.012	1000
		3	.015	1.015	-.381**	.218	.025	.140	.042	.010	1000
Parámetros estimados con MV	10	1	-.545	1.468	-1.340**	2.624**	.090**	.000	.042	.025	975
		2	-.482	1.429	-1.451**	2.906**	.097**	.000	.050	.029	992
		3	.240	.943	-1.263**	1.729**	.138**	.000	.050	.022	1000
	25	1	-.388	1.322	-.958**	1.948**	.060**	.000	.045	.014	996
		2	-.282	1.271	-1.049**	2.012**	.065**	.000	.037	.021	997
		3	.103	.885	-.625**	.781**	.043**	.000	.039	.012	1000
	50	1	-.289	1.206	-.615**	.958**	.048**	.000	.044	.013	997
		2	.072	.906	-.238**	.225	.030*	.032	.052	.009	1000
		3	.092	.966	-.473**	.623**	.046**	.000	.044	.010	999
	75	1	.022	.917	-.498**	1.249**	.059**	.000	.049	.022	999
		2	.034	.908	-.226**	.136	.030*	.031	.051	.011	996
		3	-.192	1.147	-.747**	1.396**	.045**	.000	.046	.017	996
Parámetros estimados con EAP	10	1	.256	.918	-.908**	.670**	.089**	.000	.052	.017	1000
		2	.248	.898	-1.047**	1.212**	.091**	.000	.045	.022	1000
		3	.231	.924	-1.266**	1.760**	.149**	.000	.051	.023	1000
	25	1	.244	.883	-.436**	-.046	.045**	.000	.049	.011	1000
		2	.235	.866	-.493**	-.037	.049**	.000	.042	.011	1000
		3	.226	.906	-.719**	.738**	.057**	.000	.039	.011	1000
	50	1	.197	.929	-.361**	-.075	.037**	.002	.049	.012	1000
		2	.198	.926	-.336**	.162	.026	.100	.046	.009	1000
		3	.200	.957	-.491**	.402**	.051**	.000	.050	.010	1000
	75	1	.161	.931	-.603**	1.015**	.051**	.000	.047	.016	1000
		2	.155	.925	-.307**	.090	.036**	.003	.052	.008	1000
		3	.146	.924	-.310**	-.047	.029	.050	.043	.012	1000

* p<.05; ** p<.01
[a] La presencia de N<1000 es debido a que la estimación por máxima verosimilitud marginal no converge cuando todas las respuestas del patrón son 0s o 1s.

Media de l_z

Cuando la distribución de habilidad es sesgada negativa (Tabla 3) y se emplean parámetros verdaderos, independientemente de la longitud del test y del parámetro a_j, las medias de la distribución de l_z varían en un rango de -.082, obtenido en C1 del test de 75 ítems, y .030 del test de 25 ítems en C1. Mientras que las medias infravaloradas se alejan de 0 en cantidades mínimas como -.017 de C3 del test de 25 ítems, la media sobrevalorada más próxima a 0 es .001 del test de 75 ítems también en C3.

Cuando se recurre a los parámetros estimados con MV para estudiar la distribución de l_z, la media es infravalorada en la mayoría de

Tabla 2
Estadísticos descriptivos de l_z, prueba de normalidad y FP con distribución de habilidad sesgada positiva

θ	n	C	Media	DT	Sesgo	Curtosis	M.D.	p	FP .05	FP .01	N[a]
Parámetros verdaderos	10	1	.012	.995	-.913**	1.086**	.067**	.000	.036	.020	1000
		2	-.080	1.018	-.879**	.623**	.075**	.000	.051	.016	1000
		3	-.091	1.082	-1.314**	2.705**	.099**	.000	.045	.022	1000
	25	1	.015	1.020	-.500**	-.004	.051**	.000	.048	.011	1000
		2	-.012	.963	-.481**	.265	.035**	.005	.048	.013	1000
		3	-.055	1.014	-.582**	.292	.057**	.000	.043	.010	1000
	50	1	-.020	1.046	-.412**	.258	.037**	.003	.040	.011	1000
		2	.004	.960	-.370**	.024	.050**	.000	.053	.012	1000
		3	.035	1.018	-.778**	1.600**	.057**	.000	.052	.013	1000
	75	1	-.050	1.005	-.215**	-.121	.022	.298	.053	.007	1000
		2	.012	.992	-.367**	.434**	.035**	.006	.049	.012	1000
		3	.014	1.001	-.431**	.192	.038**	.002	.043	.015	1000
Parámetros estimados con MV	10	1	-.494	1.474	-1.504**	3.324**	.096**	.000	.049	.025	980
		2	-.454	1.338	-1.393**	3.308**	.080**	.000	.043	.021	993
		3	-.404	1.439	-1.777**	4.633**	.119**	.000	.049	.027	999
	25	1	-.345	1.299	-.857**	1.232**	.069**	.000	.048	.015	989
		2	.069	.798	-.433**	.410**	.031*	.024	.055	.016	1000
		3	.093	.866	-.530**	.337*	.049**	.000	.041	.008	1000
	50	1	.051	.923	-.359**	.563**	.030*	.037	.045	.012	997
		2	.067	.899	-.362**	.050	.044**	.000	.049	.016	995
		3	.091	.939	-.671**	1.230**	.052**	.000	.052	.014	1000
	75	1	.020	.872	-.055	.252	.034**	.009	.053	.011	988
		2	.032	.898	-.299**	.619**	.027	.084	.048	.013	985
		3	.060	.934	-.398**	.347*	.034**	.008	.052	.011	991
Parámetros estimados con EAP	10	1	.240	.909	-1.151**	1.724**	.103**	.000	.044	.018	1000
		2	.243	.879	-1.010**	.966**	.103**	.000	.050	.019	1000
		3	.236	.887	-1.146**	1.349**	.129**	.000	.049	.017	1000
	25	1	.237	.891	-.551**	.218	.052**	.000	.046	.012	1000
		2	.223	.835	-.566**	.312*	.048**	.000	.042	.015	1000
		3	.207	.880	-.625**	.307*	.059**	.000	.041	.008	1000
	50	1	.197	.946	-.491**	.558**	.044**	.000	.045	.010	1000
		2	.191	.898	-.441**	.100	.044**	.000	.052	.014	1000
		3	.193	.946	-.746**	1.233**	.054**	.000	.050	.015	1000
	75	1	.182	.905	-.236**	.121	.030*	.037	.052	.010	1000
		2	.171	.911	-.379**	.612**	.031*	.024	.053	.014	1000
		3	.041	.933	-.331**	-.061	.038**	.002	.045	.010	1000

* p<.05; ** p<.01
[a] La presencia de N<1000 es debido a que la estimación por máxima verosimilitud marginal no converge cuando todas las respuestas del patrón son 0s o 1s.

las condiciones y sólo es sobrevalorada en C2 y C3 del test de 10 ítems, .176 y .207, y en C3 del test de 25 ítems, .116. El incremento tanto en el número de ítems como en el parámetro a_j provocan que los valores medios de l_z tienda al valor estandarizado; comparando los tests de 50 y 75 ítems en orden de C1 a C3 las medias son, en el primero de ellos, -.291, -.211 y -.197, y con 75 ítems, -.261, -.182 y -.166.

Tras el procedimiento EAP, las medias de l_z más cercanas al valor estándar de la distribución normal se obtienen en el test de 75 ítems, .135, .141 y .132 en C1, C2 y C3, respectivamente. La aproximación a 0 también ocurre dentro de cada longitud de test cuanto mayor es el parámetro de discriminación; por ejemplo, en el test con 25 ítems de C1 a C3 las medias son .251, .244 y .240.

Tabla 3
Estadísticos descriptivos de l_z, prueba de normalidad y FP con distribución de habilidad sesgada negativa

θ									FP		N[a]
	n	C	Media	DT	Sesgo	Curtosis	M.D.	p	.05	.01	
Parámetros verdaderos	10	1	.002	1.009	-.807**	.615**	.067**	.000	.044	.018	1000
		2	.018	1.018	-1.008**	.985**	.095**	.000	.051	.019	1000
		3	-.053	1.049	-1.107**	1.198**	.098**	.000	.048	.021	1000
	25	1	.030	.982	-.438**	.117	.036**	.003	.051	.011	1000
		2	-.055	.986	-.465**	.183	.035**	.005	.046	.012	1000
		3	-.017	.961	-.669**	.603**	.050**	.000	.044	.019	1000
	50	1	-.024	1.018	-.320**	-.007	.032*	.020	.049	.012	1000
		2	.023	.995	-.501**	.409**	.037**	.002	.044	.011	1000
		3	.006	.973	-.513**	.233	.045**	.000	.046	.015	1000
	75	1	-.082	1.048	-.378**	.135	.039**	.001	.054	.014	1000
		2	.003	.951	-.235**	-.093	.038**	.002	.047	.010	1000
		3	.001	.950	-.404**	.109	.040**	.001	.042	.011	1000
Parámetros estimados con MV	10	1	-.519	1.398	-1.266**	2.548**	.080**	.000	.041	.023	966
		2	.176	.928	-1.087**	1.820**	.109**	.000	.046	.018	999
		3	.207	.931	-1.123**	1.085**	.135**	.000	.046	.022	1000
	25	1	-.399	1.291	-.842**	1.115**	.064**	.000	.050	.019	993
		2	-.320	1.211	-.863**	1.359**	.063**	.000	.042	.014	994
		3	.116	.847	-.504**	.042	.056**	.000	.049	.012	1000
	50	1	-.291	1.241	-.633**	.343*	.056**	.000	.055	.018	993
		2	-.211	1.190	-1.018**	3.133**	.056**	.000	.036	.015	996
		3	-.197	1.147	-.799**	1.070**	.052**	.000	.044	.021	998
	75	1	-.261	1.219	-.582**	.440**	.045**	.000	.047	.012	996
		2	-.182	1.082	-.519**	.481**	.046**	.000	.045	.010	995
		3	-.166	1.099	-.674**	.724**	.069**	.000	.045	.014	996
Parámetros estimados con EAP	10	1	.246	.953	-1.002**	1.153**	.081**	.000	.048	.022	1000
		2	.243	.925	-1.144**	1.925**	.115**	.000	.049	.018	1000
		3	.241	.914	-1.153**	1.133**	.131**	.000	.052	.020	1000
	25	1	.251	.908	-.530**	.260	.058**	.000	.047	.015	1000
		2	.244	.890	-.540**	.213	.049**	.000	.046	.014	1000
		3	.240	.868	-.595**	-.032	.059**	.000	.046	.013	1000
	50	1	.198	.903	-.382**	.103	.034**	.007	.052	.011	1000
		2	.196	.938	-.600**	.787**	.049**	.000	.039	.013	1000
		3	.203	.903	-.497**	-.018	.051**	.000	.053	.012	1000
	75	1	.135	.947	-.427**	-.022	.043**	.000	.047	.014	1000
		2	.141	.912	-.369**	.129	.050**	.000	.047	.016	1000
		3	.132	.932	-.456**	.050	.041**	.001	.042	.011	1000

* p<.05; ** p<.01
[a] La presencia de N<1000 es debido a que la estimación por máxima verosimilitud marginal no converge cuando todas las respuestas del patrón son 0s o 1s.

Desviación típica de l_z

Si la distribución de habilidad es no sesgada (Tabla 1) y los parámetros empleados en el cálculo de l_z son los verdaderos, la desviación típica se aproxima bastante a 1 en todas las condiciones experimentales, siendo el valor más bajo .985 aparecido en C2 del test de 75 ítems y el valor más alto 1.061 de C3 del test de 50 ítems.

Cuando se calcula l_z con los parámetros estimados por MV, los valores más altos de dispersión se hallan en el test de 10 ítems, 1.468 en C1 y 1.429 en C2 que se corresponden a su vez con los valores más alejados de 0 de la media de l_z, -.545 en C1 y -.482 en C2. A mayor número de ítems y parámetros de discriminación, mayores son las semejanzas con la desviación típica esperadas.

Las desviaciones típicas de l_z cuando los parámetros se estiman por EAP son cercanas a las estandarizadas; el valor más próximo a 1 es .957 en C3 del test de 50 ítems y el más alejado es .866 en C2 del test con 25 ítems. Este estadístico no está afectado por la longitud del test y se aproxima más a 1 con el incremento de a_j.

Índice de simetría de l_z

Cuando la distribución de habilidad es sesgada negativa (Tabla 3) y se emplean parámetros verdaderos, las distribuciones de l_z son asimétricas negativas con p<.01 sobre todo cuanto mayor es el parámetro de discriminación; e.g., en el test de 25 ítems en C1 el índice de asimetría es -.438, en C2 es -.465 y en C3 es -.669. Este estadístico también es influenciable por la longitud del test, intentando restablecer la simetría de la curva a medida que se crece en ítems; así, con n=75 en C1 el sesgo es -.378, en C2 es -.235 y en C3 es -.404. Las distribuciones más sesgadas son las de C2 y C3 si n=10, -1.008 y -1.107, respectivamente.

El sesgo de la distribución si se utilizan los estimadores máximo-verosímiles es negativo y presenta diferencias significativas con la simetría al nivel de .01, definiendo curvas asimétricas y más cuanto menor es el número de ítems del test; con 10 ítems el sesgo es -1.266 en C1, -1.087 en C2 y -1.123 en C3, y se reduce la asimetría en el test de 75 ítems a niveles de sesgo que son en C1 -.582, en C2 -.519 y en C3 es -.674. Una excepción a esta tendencia del índice de asimetría está en C2 del test de 50 ítems, -1.018, en donde resulta una distribución tan asimétrica como las de los tests con 10 ítems.

Con parámetros estimados por EAP, las distribuciones más asimétricas negativas son las de los tests con 10 ítems, -1.002 en C1, -1.144 en C2 y -1.153 en C3, y se reduce la asimetría conforme aumenta el número de ítems, llegando a índices de sesgo -.427, -.369 y -.456 en C1, C2 y C3 si n= 75, aún siendo estadísticamente significativas. Cuanto menor es el parámetro de discriminación también son menos asimétricas las curvas, pero siempre significativas con p<.01.

Índice de curtosis de l_z

Con distribución de habilidad sesgada positiva (Tabla 2) y parámetros verdaderos, las distribuciones son leptocúrticas con p<.01 en el test de 10 ítems (1.086 en C1, .623 en C2 y 2.705 en C3), al igual que las de C3 del test de 50 ítems (1.600) y C2 del test de 75 ítems (.434). Los índices que dibujan curtosis media-baja aparecen en C1 de los tests con 25, -.004, y 75 ítems, -.121, y la curtosis media más alta está en C3 del test de 25 ítems, .292. Los tests más largos son los que muestran las distribuciones de altura media más satisfactoria.

En general, la forma de las curvas obtenidas con parámetros estimados por MV son leptocúrticas, destacando las tres condiciones del test de 10 ítems: 3.324 en C1, 3.308 en C2 y 4.633 en C3 con p<.01. En C3 de los tests de 25 y 75 ítems, .337 y .347, la significación ocurre al nivel de .05. Las distribuciones mesocúrticas de l_z son las del test de 50 ítems en C2 (.050) y C1 del test con 75 ítems (.252). El estadístico de forma tiende a 0 a mayor tamaño del test pero no parece notable el parámetro de discriminación.

Por último, con parámetros estimados por EAP, el índice de forma describe varias distribuciones leptocúrticas tanto con p<.05 como con p<.01, sin efecto aparente del parámetro de discriminación y sin trascendencia del número de ítems, con valores tan altos como los del test con 10 ítems en C1, 1.724, C2, .966 y C3, 1.349, y en C3 del test de 50 ítems, 1.233. Otras curvas son mesocúrticas: las de C1 si n= 25, el índice de curtosis es .218, C2 si n= 50 es .100, C1 y C3 del test de 75 ítems son .121 y -.061, respectivamente.

Prueba no paramétrica de Lilliefors

La prueba de normalidad pone de manifiesto la desviación de la ley normal de la distribución de l_z con p<.01 en más del 80% de las condiciones experimentales, llegando al 100% de las condiciones estudiadas cuando se han empleado parámetros estimados y la distribución de θ es sesgada negativa.

Falsos positivos

A) Distribución de habilidad no sesgada (Tabla 1): en el nivel α nominal igual a .05, el error tipo I es más consistente que al nivel .01, siendo en este nivel con parámetros verdaderos donde se producen mayores fluctuaciones en la estimación de dicho error. B) Distribución de habilidad sesgada positiva (Tabla 2): las tasas de FP más elevadas están en el nivel nominal .01 cuando n= 10, tanto con parámetros verdaderos como con parámetros estimados. En el resto de longitudes de tests dichas tasas son consistentes y muestran cierta tendencia a la infraestimación si el nivel nominal es .05 y a la sobrestimación si es .01. C) Distribución de habilidad sesgada negativa (Tabla 3): las tasas de error tipo I son más consistentes al nivel nominal .05 que al de .01, siendo a este nivel donde se sobrestima la tasa de error sobre todo en los tests de 10 ítems y con parámetros estimados por MV.

Conclusiones

En la práctica es difícil encontrar una muestra de sujetos cuya habilidad o rasgo que se va a evaluar con un test sea exactamente normal y no sesgada, pero esto, estableciendo las pertinentes limitaciones de este estudio de simulación, no parece ser un aspecto del todo relevante en lo que se refiere a la estimación del parámetro de habilidad según los resultados del estudio de recubrimiento; en cuanto a su importancia para con la distribución de l_z será analizada en las líneas siguientes.

Si se toma como referencia la prueba de normalidad de Lilliefors (1967; Marascuilo y McSweeney, 1977) la distribución de l_z no sigue la ley normal bajo las condiciones experimentales aquí planteadas salvo en contadas ocasiones, ya que existe significación estadística casi en la totalidad de las mismas. Sin embargo, estos resultados deben interpretarse con cautela porque, como toda prueba estadística, se deja afectar por el tamaño muestral; ade-

más, al igual que la prueba de Kolmogorov-Smirnov, puede rechazar la hipótesis nula de normalidad por divergencias con la distribución normal estandarizada bien en el parámetro de dificultad, en la escala de medida de los parámetros, en el índice de asimetría, o bien en el índice de curtosis; estos dos últimos podrían ser los causantes de la significación en este estudio junto con el tamaño muestral (N=1000).

Analizando las medias y desviaciones típicas para catalogar de normal la distribución de índice de medición apropiada l_z se podría confirmar que no se desvirtúa la normalidad de este estadístico y se podría recurrir a él para identificar PAR, manteniendo l_z el estatus del mejor estadístico para detectar este tipo de respuesta indeseadas. Con respecto al estadístico de tendencia central y de dispersión de l_z, los valores más cercanos a los esperados siempre se han obtenido empleando los parámetros verdaderos de la habilidad y de los ítems, así como, en el caso de la media, cuando la distribución de habilidad es centrada y no sesgada. El aumento del número de ítems y del parámetro a_j aproxima los resultados a los estándares, siendo la media infravalorada con parámetros verdaderos y estimadores máximo-verosímiles, y sobrevalorada con estimadores bayesianos, y la desviación típica infravalorada con el uso de parámetros estimados por EAP y sobrevalorada con parámetros verdaderos y estimados por MV.

No cabe lugar a dudas que la distribución de l_z es asimétrica negativa sin que se haya mostrado corrección alguna por el empleo de parámetros verdaderos o de sus estimadores ni por el grado de discriminación de los ítems. Con estimadores máximo-verosímiles aparecieron las curvas más negativamente sesgadas, junto con las descritas cuando la distribución de habilidad era centrada pero también asimétrica negativa. La asimetría de la curva es menos acusada en tests largos y con parámetros de discriminación bajos.

El índice de curtosis delinea curvas leptocúrticas sobre todo cuando l_z se ha calculado con los estimadores producto del proceso de MV. La longitud del test es un factor importante en el estadístico de forma, ya que en los tests más cortos es en los que las curvas eran más altas. No hubo un efecto definido de la distribución de habilidad sobre la altura de las curvas ni del poder discriminativo de los ítems.

Las tasas de FP se mantienen próximas a los niveles nominales, siendo más consistentes si el nivel nominal es .05. Cuando el nivel nominal es de .01 hubo tendencia a la sobrestimación de dichas tasas, sobre todo en el test más corto (10 ítems) y con parámetros estimados por MV. En definitiva, la prueba estadística de l_z para identificar PAR es conservadora y consistente en el nivel de significación nominal de .05, a pesar del sesgo y la curtosis de su distribución.

Los resultados de este estudio respecto a la distribución de l_z coinciden en su mayoría con los de Li y Olejnik (1997), aunque en desacuerdo con lo referente al sesgo ya que en su investigación la distribución de l_z era asimétrica positiva, Nering (1995, 1997), Noonan *et al.*, (1992), Reise (1995), Reise y Due (1991), y van Krimpen-Stoop y Meijer (1999, 2000), esto es, la distribución de l_z es centrada, asimétrica negativa y moderadamente leptocúrtica. Ante la dificultad de poder trabajar con los parámetros de habilidad verdaderos es conveniente recurrir a la estimación EAP de dicho parámetro y con tests de longitud próxima a 50 ítems o más, ya que con tests con menos ítems no se satisface la teoría asintótica. El sesgo de la distribución de habilidad no ha repercutido en gran medida, pero es un factor importante que se debe tener en cuenta en el momento de interpretar los resultados.

La valoración de este estudio experimental queda limitado al ámbito del formato de respuesta dicotómico, por lo que futuras líneas de investigación podrían contemplar el análisis de la medición apropiada con modelos de respuesta politómica. Otro campo a explorar es el de la relación existente entre los PAR y el funcionamiento diferencial de los ítems, ya que la causa de aquellos puede estar en la presencia de funcionamiento diferencial de los ítems cuando éstos son estimados en distintos grupos previamente igualados en un nivel de θ, y viceversa, la presencia de PAR puede ocasionar que aparezcan diferencias estadísticas en los parámetros de los ítems cuando son dichos patrones los que están desacreditando el supuesto de invarianza de los parámetros. También, evaluar la relación entre PAR y los programas de entrenamiento para mejorar las puntuaciones de los sujetos en los tests; por un lado, la ocurrencia de PAR podría modificar la valoración, tanto en sentido favorable como desfavorable, del efecto de estos programas en los sujetos a los que se les aplica; por otro, el entrenamiento para la mejora de las puntuaciones, más que alcanzar el objetivo que persigue, provoque PAR, con la repercusión que esto tiene sobre las propiedades psicométricas de los tests (Martínez-Cardeñoso, García Cueto y Muñiz, 2000; Martínez-Cardeñoso, Muñiz y García Cueto, 2000). Y, por supuesto, profundizar en las peculiaridades de un patrón de respuesta atípico en función de la variable a evaluar (rendimiento, personalidad, patologías…) y del tipo de test, e.g., lápiz y papel o TAI.

Referencias

Drasgow, F. y Guertler, E. (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. *Journal of Applied Psychology, 72*, 10-18.

Drasgow, F., Levine, M.V. y McLaughlin, M.E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59-79.

Drasgow, F., Levine, M.V. y Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

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: Latent trait test theory and computerized adaptive testing* (pp. 31-49). New York: Academic Press.

Levine, M.V. y Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology, 35*, 42-56.

Levine, M.V. y Rubin, B.D. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269-290.

Li, M.F. y Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21*, 215-231.

Lilliefors, H.W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 62*, 399-402.

Marascuilo, L.A. y McSweeney, M. (1977). *Nonparametric and distribution-free methods for social sciences*. Monterey, CA: Cole Publishing Company.

Martínez-Cardeñoso, J., García Cueto, E. y Muñiz, J. (2000). Efecto del entrenamiento sobre las propiedades psicométricas de los tests. *Psicothema, 12*, 358-362.

Martínez-Cardeñoso, J., Muñiz, J. y García Cueto, E. (2000). Mejora de las puntuaciones de los tests mediante entrenamiento. *Psicothema, 12*, 363-367.

Meijer, R.R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina y Whitney study. *Applied Psychological Measurement, 21*, 99-113.

Meijer, R.R. (1998). Consistency of test behaviour and individual difference in precision of prediction. *Journal of Occupational and Organizational Psychology, 71*, 147-160.

Meijer, R.R. y Nering, M.L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement, 21*, 321-336.

Meijer, R.R. y Sijtsma, K. (1999). *A review of methods for evaluating the fit of item score patterns on a test* (Research Report No. 99-01). Twente, The Netherlands: University of Twente, Department of Educational Measurement and Data Analysis.

Meijer, R.R. y Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.

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

Molenaar, I.W. y Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*, 75-106.

Narayanan, P. y Swaminathan, H. (1996). Identification of items that show non-uniform DIF. *Applied Psychological Measurement, 20*, 257-274.

Nering, M.L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19*, 121-129.

Nering, M.L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*, 115-127.

Noonan, B.W., Boss, M.W. y Gessaroli, M.E. (1992). The effect of test length and IRT model on the distribution and stability of three appropriateness indexes. *Applied Psychological Measurement, 16*, 345-352.

Reise, S.P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*, 213-229.

Reise, S.P. y Due, A.M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15*, 217-226.

Reise, S.P. y Flannery, Wm. P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education, 9*, 9-26.

Schmitt, N., Chan, D., Sacco, J.M., McFarland, L.A. y Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23*, 41-53.

Schmitt, N., Cortina, J.M. y Whitney, D.J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement, 17*, 143-150.

SYSTAT v. 10.0. [Computer software]. (2000). Chicago: SPSS, Inc.

van Krimpen-Stoop, E.M.L.A. y Meijer, R.R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*, 327-345.

van Krimpen-Stoop, E.M.L.A. y Meijer, R.R. (2000). Detecting person-misfit in adaptive testing using statistical process control techniques. En W.J. van der Linden y C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201-219). Boston: Kluwer-Nijhoff Publishing.