Psicothema
2000. Vol. 12, nº 3, pp. 376-382

ISSN 0214 - 9915 CODEN PSOTEG

Fuentes potenciales de sesgo en una prueba de aptitud numérica

Paula Elosua Oliden, Alicia López Jáuregui y Josu Egaña Makazaga
Universidad del País Vasco

En este trabajo se analizan dos problemas derivados de la aplicación de pruebas de aptitud numérica a una población que abarca más de un curso académico (4º y 6º de enseñanza primaria). Por un lado, el diferente grado de desarrollo cognitivo y por otro el intervalo entre la administración y la instrucción. Su interacción con la naturaleza del ítem puede ser fuente de sesgo. Para evaluar estas hipótesis, se hace uso de toda la metodología derivada del estudio del sesgo (detección del funcionamiento diferencial de los ítems y análisis de contenido), comparándose los resultados de la aplicación del χ^2 de Lord (modelo logístico de tres parámetros) y del estadístico Mantel-Haenszel. Además se pone a prueba el DIMTEST como técnica no-paramétrica para la evaluación de la unidimensionalidad esencial.

Potential sources of bias in a numerical aptitud test. In this work two problems derived from the application of numerical aptitud tests on a population which is made up of more than one academic grade (4^{th} and 6^{th} grade at primary level) are evaluated. On one hand, the different level of cognitive development and, on the other hand, the time period between administering and instruction of the tests. The interaction with the nature of the item may be the source of bias. In order to evaluate these hypotheses, all the methodology derived from the study of bias are utilisec (detection of the differential item functioning and the analysus of content). The results of the application of Lord's χ^2 (3 parameter logistic model) and the Mantel-Haenszel statistic are compared. Furthermore, the DIMTEST is put to the test as a non-parametric techniqye for the evaluation of the essential unidimensionality.

La capacidad matemática se define desde el enfoque del procesamiento de la información, como el conjunto de operaciones cognitivas, habilidades y conocimientos que son componentes de las tareas matemáticas y se analiza, en términos de los procesos cognitivos y conocimientos requeridos para su resolución. Mayer (1985) los sistematiza en 4 fases; a. *Representación.* Consiste en la traducción o conversión de los enunciados verbales a una representación interna, para lo que son necesarios tanto conocimientos acerca del lenguaje como conocimientos generales o factuales acerca de la realidad. b. *Integración* en una representación coherente. Esta fase exige un conocimiento esquemático acerca de los diferentes tipos de problemas, que permita dar a la información un significado global. c. *Planificación del problema*, para lo que se requerirán conocimientos de tipo estratégico que faciliten la elección de un modo o procedimiento adecuado para su abordaje y resolución. d. *Ejecución,* proceso que consistirá en la realización de las operaciones o algoritmos que se precisen para obtener la solución.

La evaluación de esta aptitud se lleva cabo a través de pruebas diseñadas con referentes académicos vinculados a los diseños curriculares de la enseñanza institucionalizada (Zorroza y Sánchez-Cánovas, 1995). Un análisis del contenido de estos instrumentos a la luz de la descripción de tareas expuesta, evidencia que están integrados por ítems de diferente naturaleza y complejidad en cuanto a los procesos cognitivos implicados en su resolución. Dentro de esta heterogeneidad podemos encontrar: a. Problemas de enunciado que exigen habilidades complejas de representación, interpretación, planificación y ejecución, junto a los conocimientos específicos inherentes a cada proceso. b. Items que exigen simplemente la ejecución de operaciones aritméticas, sin contenido verbal alguno, y para los cuales se precisará exclusivamente del conocimiento de los hechos aritméticos y los algoritmos de cálculo. c. Items referentes a cuestiones *teóricas* que suponen un conocimiento exclusivamente factual, *declarativo* (Anderson, 1976), en los que los requerimientos de integración y planificación son inexistentes, y los conocimientos lingüísticos necesarios en la fase de representación son mínimos. En este punto cabría apuntar que los contenidos o conocimientos factuales a que se refieren los ítems declarativos son heterogéneos en cuanto a su grado de refuerzo a lo largo del desarrollo educativo. Unos contenidos se refieren a objetivos didácticos que constituyen *ejes de fuerza* a lo largo del proceso de instrucción, mientras otros tienen un carácter marginal.

Partiendo de esta clasificación de los ítems en, ítems de enunciado, ítems aritméticos e ítems factuales o declarativos y teniendo en cuenta por un lado que los procesos cognitivos evolucionan con la edad, y por otro, que el contenido de los ítems es en muchos casos exactamente igual a los ejercicios y cuestiones que componen el material curricular, se plantean dos cuestiones en las que intentamos ahondar en este trabajo.

1. En los casos en que el grupo normativo o población destinataria de las pruebas de aptitud numérica cubre un rango de edad

Correspondencia: Paula Elosua Oliden
Facultad de Psicología
Universidad del País Vasco
20009 San Sebastián (Spain)
E-mail: pspelolp@scehu.es

que abarca más de un curso académico, la evolución con la edad de las capacidades cognitivas implicadas en la resolución de los ítems de enunciado (Riley, Greeno y Heller, 1982) puede introducir un error sistemático en el proceso de medida de la aptitud numérica, favoreciendo a los grupos de mayor edad frente a los que cursan niveles inferiores.

2. Dado que un aprendizaje exitoso requiere del almacenamiento de información en estructuras significativas relacionadas con el conocimiento previo y experiencia del sujeto, y que son más fácilmente olvidados los contenidos aislados, con ausencia de relaciones dentro de la estructura del conocimiento y el conocimiento previo (Ausubel, Novak y Hanesian, 1978), formulamos la siguiente hipótesis: la existencia de una relación positiva entre la resolución correcta del ítem y la proximidad temporal a la exposición en el aula de los contenidos a los que éste se refiere. Esta relación se verá favorecida cuando los contenidos exigidos no constituyan ejes de fuerza en el proceso educativo, y por tanto se les ha dedicado menor tiempo de instrucción.

La metodología utilizada en este trabajo se basa en el estudio del funcionamiento diferencial de los ítems, como paso previo a la evaluación del sesgo que pueden introducir en las pruebas de aptitud numérica las hipótesis planteadas. Con ello pretendemos cumplir un doble objetivo metodológico. Por un lado ampliar el campo de utilización del concepto de funcionamiento diferencial del ítem, y profundizar en las características de dos técnicas de detección (chi-cuadrado de Lord y estadístico Mantel-Haenszel), y por otro, poner a prueba un procedimiento no parámetrico para la evaluación de la unidimensionalidad esencial de datos binarios, el DIMTEST (Stout, 1987).

Método

Participantes

La muestra está formada por 356 niños con edades comprendidas entre los 9 y los 11 años que estudian en los cursos 4º (N=211) y 6º (N=145) de enseñanza primaria. De ellos 139 pertenecen a un centro de enseñanza público y los 217 restantes a un centro privado concertado de Vitoria-Gasteiz. Los datos provienen de la administración de una prueba de aptitud numérica aplicada en Mayo del curso escolar 1994-95 por una persona especialmente instruida para ello. El test pertenece a la Batería de Aptitudes Diferenciales y Generales en su versión elemental (BADYG-E) (Yuste, 1988), que cuantifica el razonamiento numérico y la aplicación de operaciones numéricas en problemas lógico-numéricos. Consta de 25 ítems de elección múltiple con 5 alternativas de respuesta. El coeficiente de fiabilidad aportado por el autor y calculado por el método de dos mitades con la corrección de Spearman-Brown es de 0,86 para 4º curso; el manual no incorpora la información correspondiente a 6º curso, ni los índices de consistencia interna para cada uno de los niveles.

Evaluación de la unidimensionalidad

El análisis de la dimensionalidad de los datos es un requisito necesario en la aplicación de un modelo de respuesta al ítem unidimensional. En nuestra investigación la existencia de un sólo factor dominante se estudia con DIMTEST (Stout, 1987; Nandakumar y Stout, 1993), procedimiento no paramétrico en el que no se hacen asunciones sobre la distribución de habilidad o sobre el tipo de función de respuesta al ítem, que evalúa la dimensionalidad esencial (d_E) de datos binarios, con resultados fructuosos (Hattie, Krakowski, Rogers y Swaminathan, 1996; Nandakumar, 1994; Nandakumar y Yu, 1996; Padilla, Pérez y González, 1999).

Sea $\tilde{U} \equiv (U_1, U_2, \ldots, U_N)$ el vector de respuestas para un test de N ítems, donde U_i denota la respuesta al ítem i. Sea θ el vector de habilidades latentes, y θ un valor particular de ese vector. $\mathbf{U}=\{U_i, i \geq 1\}$ es el banco de ítems que contiene $\tilde{\mathbf{U}}$. El banco de ítems **U** es *esencialmente independiente* con respecto a la variable latente θ si **U** satisface,

$$D_N(\theta') \equiv \frac{\sum_{1 \leq i < j \leq N} \left| Cov(U_i, U_j | \theta = \theta') \right|}{\binom{N}{2}} \rightarrow 0 \text{ cuando } N \rightarrow \infty$$

para el vector de habilidad dominante θ**'**. En contraste, la asunción de independencia local requiere que $Cov(U_i,U_j|\theta=\theta)=0$, para todo θ, donde θ expresa tanto la habilidad dominante como las habilidades menores. Es decir la independencia local exige condicionar las respuestas sobre todas las habilidades intervinientes, mientras que la asunción de independencia esencial requiere sólo condicionar sobre las habilidades dominantes. Por tanto, la independencia esencial es una asunción más débil que la independencia local. La dimensionalidad esencial (d_E) de un banco de ítems **U** es la dimensionalidad mínima necesaria para satisfacer la asunción de independencia esencial.

El procedimiento divide el test en tres subtest. 1. Un subtest dimensionalmente homogéneo, AT1, seleccionado en este trabajo con los ítems que presentan mayores saturaciones factoriales una vez sometido a un análisis de componentes principales la matriz de correlaciones tetracóricas. 2. Otro subtest AT2 similar en dificultad a AT1. 3. El subtest PT formado por el resto de ítems.

Una vez agrupados los sujetos en K niveles en función de las puntuaciones obtenidas en PT, para cada K grupo se estiman dos varianzas,

$$\hat{\sigma}_k^2 = \sum_{j=1}^{J_K} Y_j^{(k)} - \bar{Y}^{(k)2} \Big| J_k \qquad \hat{\sigma}_{U.k}^2 = \sum_{i=1}^{M} \hat{p}_i^{(k)} (1 - \hat{p}_i^{(k)})^2 \Big| M^2,$$

donde M es el número de ítems que componen cada uno de los subtests; U_{ijk} es la respuesta dada al ítem *i* por el sujeto *j* del grupo *k;* J_k es el número de sujetos en el grupo *K;* $Y_j^{(k)}$ es la proporción de respuestas correctas dadas por el sujeto *j;* $\bar{Y}^{(k)}$ es la media aritmética de la proporción de respuestas correctas del grupo *K;* $\hat{p}^i_{(k)}$ es la proporción de sujetos que responden correctamente al ítem *i* en el grupo *k.*

La diferencia estandarizada entre estas dos varianzas y sumada a través de los *K* niveles determina el valor de T_1. Siguiendo el mismo proceso con los ítems del subtest AT_2, se consigue T_2. El estadístico T de Stout viene definido por $T=(T_1-T_2)/\sqrt{2}$, que sigue una distribución normal bajo la hipótesis nula de unidimensionalidad esencial.

Funcionamiento diferencial de los ítems

La definición más general de funcionamiento diferencial del ítem podría ser la aportada por Mellenbergh (1989), según la cual dada una variable Z, y con respecto a otra variable G, el ítem *i* pre-

senta FDI, si y sólo si, se satisface la siguiente desigualdad para todos los valores *g* y *z* de las variables G y Z

$$f(X|g,z) \neq f(X|z)$$

donde X corresponde al ítem objeto de estudio; Z sería la variable medida por la prueba, y G una variable aparentemente irrelevante al objeto de medida, que normalmente suele ser nominal tal como sexo, idioma, raza, edad y cuyos valores determinan la diferenciación de los grupos de referencia (R) y focal (F).

El carácter de la variable condicionante (observada o latente) permite la clasificación de las técnicas de detección del FDI en dos conjuntos, englobados bajo los epígrafes generales de invarianza condicional observada e invarianza condicional latente (Millsap y Everson, 1993). Los procedimientos incluidos en el primer grupo definen Z como la puntuación total observada. Dentro de este apartado general, se encuadran los procedimientos chi-cuadrado (Scheuneman, 1979), el estadístico Mantel-Haenszel (Holland y Thayer 1988), la estandarización (Dorans y Kulick, 1986), modelos log-lineales (Mellenbergh, 1982) y los derivados de la regresión logística (Swaminathan y Rogers, 1990).

Dentro del segundo grupo pueden incluirse todos los procedimientos derivados de la aplicación del modelo de medida propuesto por la teoría de respuesta al ítems (TRI). La TRI postula una relación entre la variable latente Z (*habilidad latente*, θ) y la probabilidad de responder correctamente al ítem ($P_i(\theta)$), definida a través de la función de respuesta o curva característica del ítem (CCI).

En esta investigación se comparan los resultados de la aplicación de dos procedimientos de detección del FDI pertenecientes a cada uno de los grupos, el estadístico Mantel-Haenszel, perteneciente al grupo de invarianza condicional observada y el chi-cuadrado de Lord (1980) (Invarianza condicional latente).

Estadístico Mantel-Haenszel

El estadístico Mantel-Haenszel (MH) (Mantel y Haenszel, 1959) es un procedimiento simple y no-iterativo para el estudio de tablas de contingencia que Holland y Thayer (1988) aplican y desarrollan en la evaluación del funcionamiento diferencial de los ítems. Hoy día, debido a su simplicidad y eficiencia se ha convertido en el procedimiento más utilizado y punto de referencia para la evaluación de todas aquellas técnicas que tienen el mismo objetivo (Hambleton, Clauser, Mazor y Jones, 1993; Millsap y Everson, 1993).

Para algunos autores es, junto con los modelos log-lineales, un desarrollo natural de los procedimientos basados en la chi-cuadrado de Scheuneman (1979) (Kok, Mellengergh y van der Flier, 1985; Van der Flier, Mellenbergh, Adèr y Wijn, 1984), mientras que para otros (Hambleton y Rogers 1989; Holland y Thayer, 1988; Thissen y Steinberg, 1988), es equiparable a los derivados de la aplicación del modelo logístico de un parámetro.

En este procedimiento, una vez dividida la puntuación total en *K* intervalos, la hipótesis nula evalúa la razón entre la proporción de respuestas correctas e incorrectas, de modo que si es la misma en los grupos de referencia (p_{Rk}, q_{Rk}) y focal (p_{Fk} y q_{Fk}) el ítem carecería de funcionamiento diferencial,

$$H_0: \frac{p_{Rk}}{q_{Rk}} = \frac{p_{Fk}}{q_{Fk}} \quad k = 1,2,\ldots, K$$

y contrasta con el siguiente estadístico, que se ajusta a una distribución χ^2 con un grado de libertad,

$$MH\,\chi^2 = \frac{\left(\left|\sum_k A_k - \sum_k E(A_k)\right| - 1/2\right)^2}{\sum_k Var(A_k)}$$

$$E(A_k) = n_{R.}\, n_{.1} / T_k; \qquad Var(A_k) = n_{R.}\, n_{F.}\, n_{.1}\, n_{.2} / T^2_k\,(T_k - 1)$$

donde *p* y *q* son respectivamente la proporción de respuestas correctas e incorrectas; el subíndice *k* indica los niveles de habilidad; *R* y *F* representan los grupos de referencia y focal, A_k es el número de sujetos del grupo de referencia que ha contestado correctamente al ítem, y T_k el número de sujetos total en el nivel de puntuación *k* y los valores *n* corresponden a las frecuencias marginales del grupo de referencia ($n_{R.}$), del grupo focal ($n_{F.}$), de todos los sujetos que responden correctamente al ítem ($n_{.1}$), y del total de sujetos que lo erran ($n_{.2}$).

Se propone además otro estadístico, la razón de proporciones (α), con el que es posible cuantificar la magnitud y el sentido del funcionamiento diferencial. El estimador consistente y eficaz de alpha se define como

$$\hat{\alpha}_{MH} = \frac{\sum A_{kR} D_{kF} / T_k}{\sum B_{kR} C_{kF} / T_k}$$

donde A_{kR} y C_{kF} representan los sujetos de los grupos de referencia y focal que han contestado el ítem correctamente, y B_{kR} y D_{kF} son aquellos que lo han contestado de modo incorrecto en cada nivel K y grupo (R o F).

e indica el número de veces que los *odds* para el grupo de referencia son mayores que para el grupo focal. En favor de su interpretabilidad, la escala de alpha se transforma en la escala simétrica Δ, cuyo punto medio es 0, a través de la ecuación

$$\Delta_{MH} = -\frac{4}{1.7} ln\left(\hat{\alpha}_{MH}\right) = -2.35\, ln\left(\hat{\alpha}_{MH}\right)$$

En nuestro trabajo la aplicación de este procedimiento se lleva a cabo con el programa MHDIF (Fidalgo, 1994), que permite la detección del funcionamiento diferencial del ítem tanto uniforme como no uniforme (Mazor, Clauser y Hambleton, 1994) e incorpora un procedimiento de purificación del criterio en dos etapas.

Teoría de respuesta al ítem

La teoría de respuesta al ítem ofrece el marco teórico más apropiado para la detección del FDI (Mellenbergh, 1982; Shepard, Camilli y Williams, 1985). Basta igualar la variable condicionante Z con la variable latente θ para obtener la definición de la función de respuesta.

$$E(X|g,\theta) = E(X|\theta) \qquad P(X=1|g,\theta) = P(X=1|\theta)$$

Un ítem muestra funcionamiento diferencial cuando su función de respuesta depende de la pertenencia al grupo. Por el contrario, en los casos en que las funciones características sean idénticas, ex-

cepto problemas de muestreo, se considera que no existe FDI (Hambleton, Swaminathan y Rogers, 1991).

Las técnicas de detección de FDI derivadas de la TRI incluyen tres tipos de procedimientos. Aquellos que comparan en los grupos de estudio los parámetros de las curvas características del ítem (a, b, c) (Lord, 1977, 1980; Mellenbergh, 1982; Wright, Mead y Draba, 1976), los basados en el cálculo de la superficie que limitan las curvas características producidas por un ítem en dos poblaciones distintas (Linn y Harnisch. 1981; Rudner, 1977; Shepard, Camilli y Williams, 1985; Kim y Cohen, 1991; Raju, 1988, 1990), y por último los basados en la comparación de modelos (Thissen, Steinberg y Wainer, 1988, 1993). Estos últimos autores lo consideran un caso particular del procedimiento propuesto por Lord, de modo que si el ajuste de los datos al modelo que incluye parámetros diferentes para el grupo de referencia y focal es significativamente mejor que aquél en que los parámetros de los ítems son idénticos en los dos grupos, se concluye presencia de funcionamiento diferencial.

En este trabajo aplicamos el procedimiento ideado por Lord para la comparación de los parámetros de los ítems, cuya hipótesis nula para el modelo más general de tres parámetros puede formularse del siguiente modo:

$$H_o : a_{iR} = a_{iF}; b_{iR} = b_{iF}; c_{iR} = c_{iF};$$

donde el subíndice *i* denota el ítem, *R* grupo de referencia y *F* grupo focal

El estadístico utilizado para el contraste de esta hipótesis es este χ^2 que se distribuye asintóticamente con *p* grados de libertad, siendo *p* el número de parámetros comparados (Lord, 1980):

$$\chi_i^2 = v_i' \sum_i v_i$$

donde, v_i' es el vector de la diferencia entre los parámetros del ítem *i* estimado en distintas muestras, y $\sum_i$ es la matriz de varianzas-covarianzas de la diferencia entre los estimadores de los parámetros.

En la aplicación de este procedimiento de detección seguimos las pautas aconsejadas por Candell y Drasgow (1988). Una vez estimados los parámetros en cada grupo, y dada la arbitrariedad de la escala de θ, se equiparan las escalas y se estima el FDI. En una segunda fase se eliminan los ítems con FDI y se reequiparan las escalas, volviendo a detectar el FDI sobre todos los ítems. Este procedimiento se ejecuta una y otra vez hasta que en dos iteraciones consecutivas los resultados sean coincidentes.

La equiparación de las escalas se lleva a cabo por el método de la *curva característica* (Stocking y Lord, 1983) implementado en el programa EQUATE (Baker, 1994) y el análisis del funcionamiento diferencial con IRTDIF (Kim y Cohen, 1992) .

Resultados

Los estadísticos descriptivos para los cursos de 4º ($\bar{X}$= 16,65, S_x= 4,19) y 6º curso ($\bar{X}$= 16,65, S_x= 4,19), muestran una diferencia de medias significativa (t=-2,902; p=0,004). La consistencia interna se evalúa con el alpha de Cronbach (1951), que arroja los valores de 0,806 para 4º y 0,788 para 6º.

Evaluación de la unidimensionalidad

En la aplicación del DIMTEST el subtest AT1 se forma con 5 ítems que selecciona automáticamente el programa de los resultados del análisis de componentes principales ejecutado a partir de la matriz de correlaciones tetracóricas. Los dos conjuntos de datos superan el test de Wilcoxon que contrasta la hipótesis de que los

Tabla 1
Parámetros y funcionamiento diferencial de los ítems (**p<0.01)

Item	4				6				F.D.I.	
	a	b	c	χ^2	a	b	c	χ^2	χ^2 Lord	ΔMH
1	0.35	-3.58	0.115	3.1	0.30	-3.98	0.080	1.5	3.73	-0.68
2	0.45	-3.55	0.113	0.3	0.485	-3.69	0.080	0.2	0.15	0.06
3	0.60	-1.98	0.114	4.4	0.341	-2.68	0.082	3.3	9.58	1.18
4	0.49	-2.12	0.114	0.5	0.763	-2.25	0.080	0.2	2.50	-0.39
5	0.59	-2.11	0.114	0.9	0.466	-2.07	0.078	4.1	1.63	1.60
6	0.46	-1.53	0.116	5.2	0.62	-1.45	0.078	1.7	1.84	0.09
7	0.73	-2.8	0.111	3.0	0.73	-1.46	0.075	2.6	2.94	-2.55
8	0.70	-1.28	0.106	2.8	0.72	-1.46	0.074	2.9	0.70	0.15
9	0.67	-2.75	0.110	3.2	0.65	-2.28	0.082	1.6	4.90	0.58
10	0.63	-2.1	0.112	3.0	0.77	-2.08	0.076	3.0	0.41	-2.07
11	1.00	-0.93	0.113	1.4	0.88	-1.26	0.066	3.7	4.47	-0.29
12	1.42	-1.56	0.104	4.6	0.88	-1.82	0.073	3.2	7.30	0.60
13	0.53	-1.03	0.113	4.0	0.50	-1.31	0.079	1.4	1.67	-0.10
14	0.90	-0.68	0.132	0.8	0.46	-1.38	0.075	2.9	12.24**	-0.09
15	1.69	-0.87	0.134	0.9	1.09	-1.50	0.075	1.8	11.95**	-1.15
16	1.08	-0.89	0.146	1.1	0.57	-1.82	0.076	1.0	14.39**	-0.74
17	0.96	0.05	0.077	7.0	0.62	-.169	0.060	3.6	3.0371	-0.34
18	0.92	-0.18	0.115	4.6	0.79	-.231	0.058	4.6	1.2586	0.57
19	0.90	0.24	0.077	3.0	0.75	-.039	0.065	1.9	2.0761	-0.76
20	1.00	-0.76	0.127	2.9	1.11	-.286	0.081	6.2	6.5721	2.16**
21	1.70	0.15	0.053	6.3	0.76	0.514	0.068	5.5	10.3811	0.58
22	0.89	-0.07	0.106	3.7	1.41	0.082	0.077	10.8	1.8948	0.84
23	0.98	0.21	0.044	5.0	0.76	1.562	0.084	3.1	29.55**	2.84**
24	0.88	-0.21	0.107	2.4	1.08	0.708	0.036	5.1	17.65**	4.04**
25	1.12	1.21	0.082	6.2	0.89	1.385	0.056	1.2	0.8299	0.17

ítems seleccionados no sean excesivamente fáciles. Para 4º y 6º los valores de *p* son correlativamente, p=0,06 y p=0,227. La aplicación de este procedimiento en su versión conservadora produce los valores de T=0,5736; p≤0,283 (T_1=-0,129; T_2=-0,9403) y T=-0,4406 p≤0,6702 (T_1=-0,9605; T_2=-0,3373) para los cursos de 4º y 6º respectivamente. Estos índices alcanzan los valores T=0,7328; p≤0,2318 (T_1=-0,2313; T_2=-1,2676) y T=-0,6082; p≤0,7284 (T_1=-1,3500; T_2=-0,4898) en su versión más potente. Puede verse que en los dos casos se acepta la hipótesis contrastada de unidimensionalidad esencial.

Estimación de los parámetros

El modelo seleccionado para la estimación de los parámetros es el modelo más general, el logístico de tres parámetros, que no exige asunciones sobre los parámetros de discriminación o azar. El procedimiento de estimación utilizado es el marginal por máxima verosimilitud implementado en BILOG (Mislevy y Bock, 1990). Los parámetros obtenidos en cada una de las muestras y lo índices de ajuste se muestran en la tabla 1.

Funcionamiento diferencial de los ítems

Aplicadas las técnicas de detección del funcionamiento diferencial de los ítems los resultados pueden verse en la tabla 1, donde los asteriscos corresponden a índices de FDI significativos (p<0,01). Bajo el epígrafe χ^2 aparecen los valores de este estadístico, tras un proceso iterativo de purificación del criterio que converge en dos etapas y cuyas constantes de equiparación son, A_1= 0,9010 - K_1= -0,2889 y A_2= 0,8058 - K_2= -0,3851. La columna Δ_{MH} muestra los índices delta tras un proceso bietápico de purificación de la puntuación observada. Los valores positivos de Δ_{MH} indican funcionamiento diferencial a favor del grupo focal, que en nuestro caso es el formado por los sujetos de 4º. Una vez aplicada la corrección propuesta por Mazor, Clauser y Hambleton (1994) para la detección del funcionamiento diferencial no uniforme, los índices que se obtienen son significativos para los ítems 20 y 24 en el grupo de puntuación inferior (Δ_{MH20}= 2,81 y Δ_{MH24} =7,55). Entre los sujetos que obtienen puntuaciones más elevadas presentan funcionamiento diferencial los ítems 23 y 24 para los que el valor Δ_{MH} es de 2,84. Es de reseñar que estos 3 ítems 20, 23 y 24 han sido también detectados en el análisis del funcionamiento diferencial uniforme, y es de destacar el hecho de que los tres favorecen al grupo de 4º.

El estadístico Mantel-Haenszel cataloga 3 ítems con funcionamiento diferencial (20, 23 y 24) mientras que el chi-cuadrado de Lord detecta 5 (14, 15, 16, 23 y 24). El número de coincidencias en la detección de ítems con funcionamiento diferencial se reduce a dos, los ítems 23 y 24.

Este nivel de acuerdo entre procedimientos se incrementaría si el modelo de respuesta al ítem utilizado en la estimación de parámetros fuera el de dos. La utilización del modelo logístico de dos parámetros en la detección del FDI que tras el proceso iterativo aconsejado por Candell y Drasgow (1988) converge en dos etapas (A_1= 0,8927 - K_1= -0,3548 y A_2= 0,8845 - K_2= -0,5373) ofrece valores significativos en los ítems 20, 23 y 24 (χ^2_{20}= 16,32; p<0,01; χ^2_{23}= 14,50; p<0,01; χ^2_{24}= 49,17; p<0,01). Este hecho evidencia un alto nivel de concordancia entre ambos procedimientos atestiguada en otras investigaciones (Hambleton y Rogers, 1989; Raju, Drasgow y Slinde, 1993; Elosua, López y Egaña, en prensa) que se ve aminorada con la incorporación del parámetro de pseudoazar al modelo, circunstancia ésta encontrada en trabajos anteriores (Elosua, López, Egaña, Artamendi y Yenes, en prensa).

Análisis de contenido

Una vez detectados los ítems que presentan funcionamiento diferencial, es necesario dar un paso más en la búsqueda de sus causas. Para ello intentamos clasificar los ítems de la prueba analizada en función de los procesos cognitivos y conocimientos que demandan.

– Existen 11 ítems sin contenido verbal, en los que se exige la ejecución de una de las cuatro operaciones aritméticas básicas. Según el modelo propuesto por Mayer demandarían únicamente el proceso de ejecución y los correspondientes conocimientos algorítmicos (ítems 1-2-3-4-5-6-8-10-11-17-19).

– Doce ítems que implican la resolución de problemas de enunciado. Ocho de los cuales requieren la realización de alguna o algunas de las operaciones aritméticas básicas (ítems 7-9-13-14-15-16-18), mientras que para la resolución de los restantes ítems son precisos de modo adicional conocimientos factuales de equivalencias de fracciones (ítem 21), magnitudes (ítem 23) y geometría (ítems 22-25).

– Dos ítems que requieren exclusivamente de conocimientos factuales de tipo *declarativo* (equivalencia de magnitudes y escritura de números romanos) en los que no es preciso ejecutar operación aritmética alguna (ítems 20-24).

Aunando los resultados estadísticos en la detección del FDI y el análisis de contenido, podemos apreciar: 1. Ausencia de FDI en todo el bloque de ítems correspondientes a operaciones aritméticas. 2. Presencia de funcionamiento diferencial a favor del curso superior en tres de los problemas de enunciado (ítems 14-15-16) y a favor del inferior en un problema de enunciado con conocimiento factual (ítem 23). 3. Funcionamiento diferencial que favorece al grupo inferior en los dos ítems declarativos (ítems 20 y 24).

Conclusiones

Expuestos los resultados obtenidos, estamos en situación de mostrar las conclusiones de carácter específico a las que hemos llegado en función de las hipótesis planteadas:

– Los ítems que exigen del sujeto la ejecución de una operación básica, en los que la carga verbal se limita a instrucciones como «multiplica» o «divide», no presentan funcionamiento diferencial. En ninguno de los 11 ítems analizados aparece tal condición. Esto no significa que no existan diferencias en la ejecución de estos ítems en los dos grupos analizados, sino que la probabilidad de respuesta correcta a éste tipo de ítems en sujetos con el mismo nivel de habilidad es independiente de su pertenencia a 4º o 6º curso.

– En el grupo de los ítems declarativos, si nos limitamos al ítem detectado simultáneamente por el chi-cuadrado de Lord y por el estadístico Mantel-Haenszel (ítem 24), podemos observar que favorece a los alumnos de 4º curso. El examen de la distribución de contenidos de los textos de matemáticas utilizados en la enseñanza primaria a lo largo de los cursos permite constatar el hecho de que es precisamente en este curso en el que se instruye a los niños en la escritura de números romanos, observándose la ausencia de referencia a dichos contenidos en niveles superiores. El menor lapso de tiempo transcurrido entre la instrucción y la ejecución de la

prueba, unido a una falta de refuerzo en el conocimiento, podría explicar esta circunstancia.

– En la categoría de ítems con problemas de enunciado son cuatro los ítems que presentan funcionamiento diferencial. De ellos, tres favorecen a los sujetos de mayor edad y uno les perjudica. Los tres primeros no exigen conocimientos factuales, mientras que el que favorece a los menores exige conocimiento acerca de la equivalencia entre horas y minutos.

Esta clase de ítems pone en juego un amplio repertorio de capacidades y conocimientos, cuya integración es necesaria para su correcta resolución. Las diferencias interpersonales e intergrupales, en nuestro caso 4º frente a 6º, pueden darse en todos y en cada uno de los procesos exigidos. Podríamos mencionar en este punto, que entre otros, el conocimiento lingüístico se desarrolla con la edad (Greeno, 1980), lo que implica un incremento en la complejidad del conocimiento conceptual requerido para la comprensión de las situaciones descritas en los problemas de enunciado (Riley, Greeno y Heller, 1982). El hecho de que en una de las habilidades implicadas en la resolución de los ítems de enunciado la distribución condicional de los grupos pueda variar, podría ser causa del funcionamiento diferencial y fuente potencial de sesgo.

– Existen otros dos ítems, uno de enunciado con necesidad de conocimiento específico (correspondencia entre horas y minutos) y otro declarativo (correspondencia entre centímetros y metros) que muestran funcionamiento diferencial a favor del grupo 4º. Siendo ambos conocimientos reforzados a lo largo de todo el proceso educativo tal vez sea la proximidad de exposición de contenidos lo que explique esta circunstancia.

Por todo ello diríamos que existen potenciales fuentes de sesgo a tener en cuenta cuando una misma prueba es aplicada a sujetos de diferentes cursos, y que estas están relacionadas con la naturaleza y contenido de los ítems. Si bien existe un tipo de problemas complejos para los cuales se requiere una realización experta y estos favorecen a los sujetos de más edad, los ítems relativos a conocimientos factuales estarán sujetos a la influencia que sobre la retención ejerza el tiempo transcurrido entre la instrucción y la ejecución de la prueba.

Si bien somos conscientes del carácter exploratorio de este análisis y de la conveniencia del diseño de estudios ad-hoc para obtener evidencias más concluyentes, creemos que el presente trabajo supone una llamada de atención acerca de cuestiones no contempladas habitualmente por los constructores de tests.

Con carácter más general y con relación a los estudios de sesgo apuntaríamos que si aceptamos la definición de sesgo como error sistemático que distorsiona el significado de las puntuaciones y que está causado por la intervención de habilidades espurias junto a la habilidad principal (Ackerman, 1992; Mellenbergh, 1989; Shealy y Stout, 1993), podemos incluir su evaluación dentro del análisis de la validez. Desde esta visión integradora, el estudio del sesgo al igual que el de la validez se convierte en un proceso continuo en el que se recogen evidencias estadísticas y lógicas, que se complementan y refuerzan (Hambleton, Clauser, Mazor y Jones, 1993; Scheuneman, 1987) para la confirmación de las hipótesis postuladas respecto a la variable medida, y para la justificación de las inferencias basadas en las puntuaciones obtenidas (Cronbach, 1971).

Los resultados obtenidos en este trabajo apoyan firmemente la inclusión de este tipo de estudio en el proceso de construcción de tests. La integración del estudio del sesgo en la evaluación de la validez, además de permitir ahondar en el conocimiento del constructo, es un medio de garantizar la equidad del proceso de medida con ítems que siendo relevantes salvaguarden las condiciones mencionadas por Title (1982) de imparcialidad de contenidos y representación positiva de grupos minoritarios. Características éstas que pueden preservarse desde el propio estadio de redacción de los ítems, con un análisis que tenga en cuenta tanto la terminología utilizada como la familiaridad de contenidos.

Referencias

Ackerman, T.A.(1992). Didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

Anderson, J. R. (1976). *Language, memory, and thought*. Hillsdale, NJ: Erlbaum.

Ausubel, David P., Joseph D. Novak, and Helen Hanesian . (1978) . *Educational psychology: A cognitive view*. New York: Holt, Rinehart and Winston, Inc . (Trad. Cast. Psicología educativa: un punto de vista cognoscitivo. México: Trillas, 1983)

Baker, F.B. (1994) EQUATE2: Computer program for equating two metrics in item response theory [Computer program]. Madison: University of Wisconsin, Laboratory of Experimental Design

Candell, G.L. y Drasgow, F.(1988): An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied psychological measurement*, 12(3), 253-260.

Cronbach, L.J.(1951). Coefficient Alpha and the Interntal Structure of Tests, *Psychometrika, 16*, 297-334.

Cronbach, L.J.(1971). Test validation. In R.L. Thorndike(Ed.), *Educational Measurement*(pp.443-507). Washington, DC: American Council on Education.

Dorans, N.J. y Kulick, E.(1986). Demonstrating the utility of the Standarization Approach to assessing unexpected Differential Item Performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*(4), 355-368.

Elosua, P. López, A y Egaña, J. (2000) Idioma de aplicación y rendimiento en una prueba de comprensión verbal. Psicothema 12(2), 201-206.

Elosua, P., López, A., Egaña, J., Artamendi, J. y Yenes, F. (Revista de Metodología de las Ciencias del Comportamiento). Funcionamiento diferencial de los ítems en la aplicación de pruebas psicológicas en entornos bilingües (En prensa).

Fidalgo, A.M.(1994). MHDIF: A computer program for detecting uniform and nouniform differentzial item functioning with the Mantel-Haenszel procedure.[Computer program] Dpto. Psicología, Universidad de Oviedo.

Greeno, J.G. (1980) Some examples of cognitive task analysis with instructional implications, en R.E.Snow, P.Federico y W.E: Montague (Eds.) *Aptitude, Learning an Instruction*. Hillsdale,N.J.Erlbaum

Hambleton, R.K., Clauser, B.E., Mazor, K.M. y Jones, R.W. (1993) Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*(1), 1-18.

Hambleton, R.K. y Rogers, H.J.(1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel Methods. *Applied Measurement in Education, 2*(4), 313-334.

Hambleton, R.K., Swaminathan, H. y Rogers, H.J.(1991). *Fundamentals of Item Response Theory.* Newbury Park: SAGE publications.

Hattie, J., Krakowski, K., Rogers, H.J. y Swaminathan, H.(1996) An assessment of Stout's index of essential unidimensionality. *Applied psychological measurement, 20*(1), 1-14.

Holland, P.W. y Thayer, D.T.(1988). Differential Item Performance and the Mantel-Haenszel procedure. En H. Wainer y H.J. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Kim, S.H. y Cohen, A.S.(1991). A comparison of two area measures for detecting Differential Item Functioning. *Applied Psychological Measurement, 15*(3), 269-278.

Kim S.H. y Cohen, A.S. (1992). IRTDIF: A computer program for IRT differential item functioning analysis [Computer Program] University of Wisconsin-Madison.

Kok, F.G., Mellenbergh, G.J. y Van der Flier, H.(1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement, 2*, 295-303.

Linn, R.L. y Harnisch, D.L.(1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*(2), 109-118.

Lord, F.M.(1977). A study of item bias, using item characteristic curve theory. En Y.H. Poortinga(Ed.), *Basic problems Cross-Cultural Psychology* (pp.19-29).Amsterdam: Swets y Zeitlinger.

Lord, F.M.(1980). *Applications of Item Response Theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Mayer, R.E. (1985). Capacidad matemática. En R.J. Sternberg, (de.) *Human abilities. An information processing approach.* New York: Freeman and company. (Trad. Cast. Las capacidades humanas. Un enfoque desde el procesamiento de la información. Barcelona. Labor, 1986)

Mantel, N. y Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute, 22,* 719-748.

Mazor, K.M., Clauser, P.E. y Hambleton, R.K. (1994). Identification of nonuniform Differential Item Functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54,* 284-291.

Mellenbergh, G.J.(1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*(2), 105-118.

Mellenbergh, G.J.(1989). Item bias and Item Response Theory. *International Journal of Educational Research, 13,* 127-143.

Millsap, R.E. y Everson, H.T.(1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297-334.

Mislevy, R.J. y Bock, R.D.(1990). BILOG·3: Item analysis and test scoring with binary logistic models.[Computer program]. Mooresville, IN: Scientific software.

Nandakumar, R. (1994) Assessing dimensionality pf a set of item responses. Comparison of different approaches. *Journal of educational measurement, 31,* 17-35.

Nandakumar, R. y Stout, W. (1993) Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of educational statistics, 18,*41-68

Nandakumar, R. y Yu, F. (1996) Empirical validation of DIMTEST on nonnormal ability distributions. *Journal of educational measurement, 33,* 355-368.

Padilla, J.L., Pérez, C. y González, A. (1999) Efecto de la instrucción sobre la dimensionalidad del test. *Psicothema, 11*(1), 183-193.

Raju, N.S.(1988). The area between two item characteristic curves. *Psychometrika,53*(4), 495-502.

Raju, N.S.(1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207.

Raju, N.S., Drasgow, F. y Slinde, J.A.(1993). An empirical comparison of the area methods, Lord's Chi-square test and Mantel-Haenszel technique for assessing Differential Item Functioning. *Educational and Psychological Measurement, 53*(2), 301-314.

Resnick, L.B. y Ford, W.W.(1981) *The psychology of mathematics for instruction,* Hillsdale, N.J. Erlbaum

Riley, M., Greeno, J.G. y Heller, J. (1982) The development of children's problem solving ability in arithmetic, en H.Ginsburg(ed.) *The development of mathematical thinking,* Nueva York:Academic Press.

Rudner, L.M.(1977, April). *An approach to biased item identification using latent trait measurement theory.* Paper presented at the Annual Meeting of The American Educational Research Association, New York.

Shealy, R. y Stout, W.(1993). An Item Response Theory model of test bias and Differential Test Functioning. In W.P. Holland y H. Wainer(Eds.), *Differential Item Functioning*(pp.197-240). Hillsadale, NJ: Lawrence Erlbaum.

Shepard, L.A., Camilli, G. y Williams, D.M.(1985). Validity of aproximation techniques for detecting item bias. *Journal of Educational Measurement, 22*(2), 77-105.

Scheuneman, J.D.(1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16*(3), 143-152.

Scheuneman, J.D.(1987). An experimental exploratory study of causes of bias in test items. *Journal of Educational Measurement, 24*(1), 97-118.

Stocking, M.L. y Lord, F.M. (1983) Developing a common metric in item response theory. *Applied psychological measurement, 7*(2), 201.210.

Stout, W.F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589-617.

Swaminathan, H. y Rogers, H.J.(1990). Detecting Differential Item Functioning using logistic regresion procedures. *Journal of Educational measurement, 27*(4), 361-370.

Title, C.K.(1982). Use of judgmental methods in item bias studies. En R.A.Berk (Ed.) *Handbook of methods for detecting test bias* (pp.31-63). Baltimore, MD:The John Hopkins University Press.

Thissen, D. y Steinberg, L.(1988). Data analysis using Item Response Theory. *Psychological Bulletin, 104*(3), 385-395.

Thissen, D., Steinberg, L. y Wainer, H.(1988). Usee of item response theory in the study of group differences in trace lines. En H.Wainer y H.I. Braun (Eds.) *Test validity* (pág. 147-169). Hillsdale, NJ: LEA

Thissen, D., Steinberg, L. y Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. En W.P. Holland y H.Wainer (Eds.) *Differential item functioning* (pág. 67-113). Huillsdale, NJ: LEA

Van der Flier, H., Mellenbergh, G.J., Adèr, H.J. y Wijn, M.(1984) An iterative item bias detection method. *Journal of educational measurement*, 21(2), 131-145.

Wright, B.D., Mead, R. y Draba, R.(1976). *Detecting and correcting item bias with a logistic response model.*(Research memorandum, Nº22). Chicago, IL: University of Chicago, Statistical lab., Departament of Education.

Yuste, C. (1988). BADYG-E. Madrid. Ciencias de la educación preescolar y especial.

Zorroza, J. y Sánchez Cánovas, J.(1995).Los componentes cognitivos de la capacidad matemática: Representación mental, esquemas estrategias y algoritmos. *Psicológica,16,* 305-320.

Aceptado el 7 de febrero de 2000