Rev. de Psicol. Gral. y Aplic., 2002, *55* (3), 349-362

ANÁLISIS DE CONTENIDO Y FUNCIONAMIENTO DIFERENCIAL EN UNA PRUEBA DE APTITUD NUMÉRICA[1]

ALICIA LÓPEZ JÁUREGUI Y PAULA ELOSÚA OLIDEN
Dpto. Psicología Social y Metodología de las Ciencias del Comportamiento
Fac. de Psicología, Universidad del País Vasco

Resumen

La aplicación de una misma prueba de aptitud a sujetos pertenecientes a diferentes cursos puede generar dos fuentes de variabilidad que podrían comportar la ausencia de equidad en el proceso de medición; por un lado la evolución de los componentes cognitivos implicados en la resolución de problemas y por otro, los diferentes intervalos de tiempo transcurridos entre la instrucción en el aula y la ejecución de la prueba. En este trabajo se analizan los resultados de la administración de una prueba de aptitud numérica a sujetos de 4º 5º y 6º curso de Enseñanza Primaria. La evaluación de las fuentes de sesgo se lleva a cabo a través del estudio del funcionamiento diferencial de los ítems (FDI) (c^2 de Lord) y un análisis posterior de contenido. Los resultados encontrados apoyan las hipótesis planteadas

Palabras clave: Validez de contenido, Funcionamiento Diferencial del Item, Sesgo.

Abstract

The application of the same test to subjects from different grades can generate two sources of variability which could hamper the equality in the testing process. The first source is the evolution of the cognitive components used in problem solving, and the second source is the different time lapses between classroom instruction and the carrying out of the test. This paper shows the analysis of the results from a numeric aptitude test to subjects in 4th, 5th and 6th grade in Elementary School. The evaluation of the sources of bias is done through the study of the differential item functioning (DFI) (Lord's c^2) and a subsequent content analysis. The results support the original hypothesis.

Key words: Content validity, Differential Item Functioning, Bias.

INTRODUCCIÓN

La aptitud numérica o matemática se ha evaluado desde la psicología cognitiva a través de pruebas diseñadas con referentes académicos vinculados a los diseños curriculares de la enseñanza institucionalizada (Zorroza y Sanchez-Cánovas, 1995), asumiendo así una relación estrecha entre el conocimiento y la resolución correcta del problema planteado (Resnick y Ford, 1981).

Sin embargo cuando el grupo normativo, o la población destinataria de las pruebas de medición, cubre un rango de edad que abarca más de un curso académico, se plantean nuevas fuentes de variabilidad que podrían suponer una amenaza para la validez. Por un lado la evolución de los componentes cognitivos implicados en la resolución de problemas y por otro, la secuenciación didáctica de contenidos. Ambos aspectos pueden incidir en los rendimientos observados.

Ahora bien, el análisis de los ítems que componen las pruebas de aptitud numérica estándar evidencia una gran heterogeneidad en función de los procesos cognitivos y tipos de conocimiento requeridos. La naturaleza y complejidad de los ítems constituye un condicionante de los componentes cognitivos puestos en juego para su resolución.

[1] Este trabajo ha sido financiado por la Universidad del País Vasco. 1/UPV/EHU 00109.231-HA-7852/2000

Desde la psicología cognitiva y en el marco de la teoría del procesamiento de la información centrada en el análisis de tareas y en la subdivisión de componentes, Mayer (1982) propone cuatro procesos y tipos de conocimiento básicos implicados en la resolución de problemas matemáticos.

1.- La *representación del problema* consiste en la traducción o conversión de los enunciados verbales a una representación interna para lo cual son necesarios tanto conocimientos acerca del lenguaje utilizado como *conocimientos generales o factuales* acerca de la realidad.
2.- La *integración* de la información en una representación coherente para lo cual será necesario poseer un *conocimiento esquemático* acerca de los diferentes tipos de problemas que permita dar a la información un significado global.
3.- La fase de *planificación* para la cual son necesarios *conocimientos de tipo estratégico* que faciliten la elección de un curso de acción adecuado para el abordaje y resolución del problema.
4.- El proceso de *ejecución* que consistirá en la realización de las *operaciones o algoritmos* que se precisen para obtener la solución.

Si se analizan las pruebas de aptitud numérica a la luz del análisis de tareas expuesto, se puede observar que junto a los problemas de enunciado que exigen los cuatro procesos anteriores, coexisten, por un lado, una categoría de ítems sin contenido verbal alguno, para los cuales se precisará exclusivamente del conocimiento de los hechos aritméticos y los algoritmos de cálculo y por otro, ítems acerca de cuestiones "teóricas" que suponen un conocimiento exclusivamente factual "declarativo" (Anderson, 1976). Para estas dos últimas clases de ítems los requerimientos de integración y planificación son inexistentes.

Teniendo en cuenta la anterior clasificación, pretendemos analizar la interacción del tipo de ítem con dos variables susceptibles de introducir un error sistemático en el rendimiento; por un lado la edad y el consecuente desarrollo cognitivo, y por otro, el efecto de la distribución de contenidos a lo largo de los cursos que generará intervalos de tiempo diferentes entre la instrucción en el aula y la ejecución de la prueba de aptitud. En concreto postulamos que la evolución con la edad de las capacidades cognitivas conducirá a una ventaja sistemática de los grupos de mayor edad frente a los que cursan niveles inferiores en aquellos ítems que ponen en juego un mayor número de procesos y conocimientos.

La segunda cuestión que planteamos en este trabajo se relaciona con la distribución de contenidos a lo largo de los cursos. Las pruebas de aptitud numérica destinadas a niños en edad escolar se componen de ítems similares, si no idénticos al material curricular. En esas circunstancias postulamos la existencia de una relación entre la resolución correcta del ítem y la proximidad temporal a la exposición en el aula de los contenidos a los que se refiere este ítem. Ahora bien existe heterogeneidad en cuanto al grado de refuerzo de los contenidos a lo largo de los cursos. Unos contenidos constituyen "ejes de fuerza" a lo largo de la instrucción siendo otros de carácter accesorio o marginal. Un aprendizaje exitoso requiere que los contenidos puedan incorporarse a las estructuras de conocimiento y experiencia del sujeto mientras que los contenidos con ausencia de relaciones dentro de la estructura de conocimiento y el conocimiento previo serán fácilmente olvidados (Ausubel, Novak y Hanesian, 1978). En consecuencia cabría esperar que los ítems referentes a contenidos aislados a los que se ha dedicado un menor tiempo de instrucción e insuficientemente consolidados se vean favorecidos en mayor grado por la proximidad en la exposición.

En síntesis, el objetivo sustantivo de este trabajo consiste en obtener evidencias que permitan determinar la existencia de las dos fuentes de sesgo expuestas Para ello haremos uso de la metodología aportada por el estudio del sesgo y derivada del desarrollo de la teoría de respuesta al ítem (TRI), profundizando en las posibles causas del mismo a través de la clasificación y el examen del contenido de los ítems que componen la prueba.

Método

Participantes

La muestra está formada por 542 niños con edades comprendidas entre los 9 y los 11 años que estudian en los cursos 4º (N=211), 5º (N=186) y 6º (N=145) de enseñanza primaria en un centro público (N=229) y uno concertado (N=313) de Vitoria-Gasteiz. Los cuestionarios se administran en Mayo del curso 1994-95 por una persona especialmente instruida para ello.

Los datos corresponden a la prueba de Aptitud Numérica de la Batería de Aptitudes Diferenciales y Generales en su versión Elementeal (BADYG-E) (Yuste, 1988). Consta de 25 ítems de elección múltiple con 5 alternativas de respuesta. Los coeficientes de fiabilidad aportados por el autor y calculados por el método de dos mitades con la corrección de Spearman-Brown arrojan los valores de 0,86 para 4º y 0,88 para 5º. El manual no incorpora información correspondiente a 6º curso, ni los índices de consistencia interna para cada uno de los niveles.

Teoría de respuesta al ítem

La teoría de respuesta al ítem (TRI) es un nuevo enfoque en la teoría de los tests que pretende superar las limitaciones de la Teoría Clásica, y cuyos principales objetivos son proporcionar mediciones que no estén en función del instrumento utilizado, y disponer de instrumentos de medida cuyas propiedades no dependan de los objetos medidos. Tiene por finalidad la obtención de mediciones invariantes respecto de las pruebas utilizadas y de los sujetos implicados (Muñiz, 1997).

Los modelos de TRI asumen la existencia de una relación entre la variable que miden los ítems y la probabilidad de responderlos correctamente. Esta relación adopta la forma de una función matemática denominada Función de Respuesta o Curva Característica del Item (CCI), que constituye el eje central de la teoría (Figura 1).

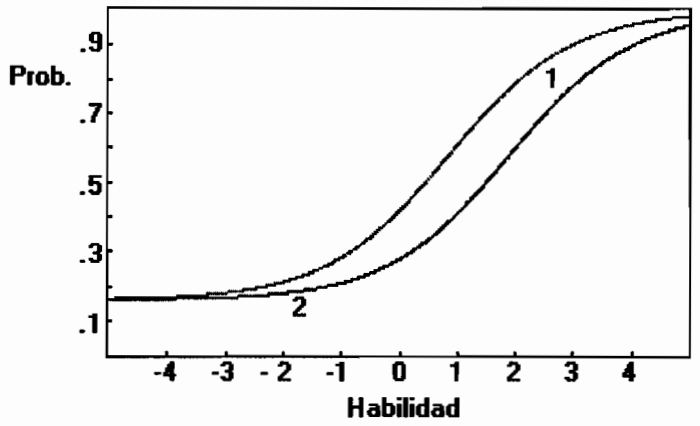
Según se adopte para la CCI una función matemática u otra y según se definan para los ítems uno, dos o tres parámetros se generarán diferentes modelos. Siendo tres los parámetros que pueden definir a los ítems (dificultad, b_i, discriminación, a_i y pseudo-azar, c_i) los modelos más utilizados son los logísticos de uno, dos y tres parámetros y los modelos de ojiva normal de uno, dos y tres parámetros. Habitualmente se opta por los primeros debido a su tratabilidad matemática.

En el modelo de un parámetro se asume que los parámetros de discriminación (a_i), y pseudo-azar (c_i) son constantes para todos los ítems. En el modelo de dos parámetros se asumen iguales parámetros de pseudo-azar (c_i). El modelo de tres parámetros, es el más general, y no se realizan asunciones sobre ninguno de los parámetros. En este último caso, la función de respuesta o curva característica del ítem vendría definida por la siguiente ecuación:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

donde θ, valores de la variable medida
$P_i(\theta)$, probabilidad de acertar al ítem i para una valor θ
b_i, parámetro de dificultad del ítem *i*
a_i, parámetro de discriminación del ítem *i*
c_i ,el valor de $P_i(\theta)$ cuando $\theta=-\infty$
e base de los logaritmos neperianos (2,72)
D, Constante. Cuando toma el valor 1,7 la función logística se aproxima a la normal.

Figura 1. Curvas características de dos ítems



Cuando las CCIs para el mismo ítem estimadas separadamente en dos grupos diferentes (grupo de referencia y grupo focal) son iguales, salvo errores de muestreo, existe equivalencia métrica, es decir, la relación entre la variable medida y el ítem es la misma en los dos grupos (Drasgow, 1987). En caso contrario estaríamos ante un funcionamiento diferencial del ítem (Hambleton y Swaminathan, 1985; Lord, 1980). En estas situaciones, la probabilidad de respuesta correcta a un ítem entre sujetos con el mismo nivel de habilidad es diferente en función del grupo de pertenencia. Esto significaría que en uno de los grupos la probabilidad de respuesta correcta no depende únicamente de la variable medida. Es decir, junto a la variable principal actuarían variables espurias que contaminarían el proceso de medición, perjudicando sistemáticamente a los miembros de un grupo.

El concepto de funcionamiento diferencial del ítem permite así la evaluación de la equivalencia métrica de los ítems entre grupos y ha sido utilizado para la adaptación de tests (Hulin y Mayer, 1986; Bontempo, 1993; Ellis, Becker y Kimmel, 1993; Elosúa y López, 1999) estudios transculturales (Ellis y Kimmel, 1992; Lord, 1977), búsqueda de sesgo contra grupos definidos en función del sexo (Gómez y Navas, 1998), minorías lingüísticas (Elosúa, López y Egaña, en prensa), o influencia de la historia instruccional sobre el rendimiento (Linn, Levine, Hastings y Wardrop, 1981; Padilla, Pérez y González, 1998).

Detección del Funcionamiento diferencial de los ítems

Dentro del marco teórico ofrecido por la TRI disponemos de varios procedimientos estadísticos para la comparación de las funciones de respuesta obtenidas en dos grupos. Unos comparan los parámetros que caracterizan a los ítems (*b, a,* y *c),* analizando su igualdad o desigualdad (Lord, 1977, 1980; Wright, Mead y Draba, 1976). Otros se basan en el cálculo de la superficie que limita las dos 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).

En este trabajo la detección del FDI se lleva a cabo comparando los parámetros de los ítems obtenidos en distintas muestras. Para ello la hipótesis nula para el modelo más general sería:

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

donde el subíndice i representa al ítem

Para la comprobación de esta hipótesis utilizamos el estadístico χ^2 (Lord, 1980):

$$\chi_i^2 = v_i' \sum_i^{-1} v_i$$

donde, v_i' es el vector de la diferencia entre los parámetros del ítem *i* estimado en distintas muestras, y $\sum_i^{-1}$ es la inversa de 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). En una primera etapa se estiman los parámetros de los ítems de modo independiente en cada uno de los grupos a analizar. Dada la indeterminación de la escala q, es decir, dado que el origen y la unidad de la escala q son arbitrarios, es necesario situar los parámetros obtenidos en dos muestras en una misma escala antes de proceder a su comparación, proceso que recibe el nombre de equiparación. Una vez equiparadas las escalas es posible estimar el funcionamiento diferencial de los ítems. En un segundo estadio del análisis y con el objetivo de mejorar el proceso de estimación, se eliminan los ítems con FDI y se vuelven a equiparar las escalas con aquellos ítems que no han mostrado funcionamiento diferencial, 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.

Sesgo

Una vez utilizados los métodos estadísticos disponibles para la detección del FDI y debido a que los índices estadísticos no son por sí mismos pruebas de sesgo, es necesario dar un paso más y buscar las razones de los resultados obtenidos. En esta fase es necesario poner en juego análisis lógicos o experimentales que determinen las causas del FDI y concluyan en su caso presencia de sesgo, lo que se interpretaría como falta de validez de la prueba.

Esto significa que dentro del análisis de la validez de todo instrumento de medición psicopedagógico es preciso estudiar el posible sesgo contra determinados grupos que pueden definirse en función de la edad, sexo, grupo cultural, minoría étnica. El análisis ha de comprender una fase exploratoria de detección del FDI, y continuar con una fase confirmatoria de evaluación del sesgo.

Este trabajo se inscribe en esta metodología que comprende el sesgo y por ende la validez como conceptos integradores, y que precisan la recogida continua de evidencias para profundizar en el conocimiento de la variable medida, asegurando de este modo la validez de la prueba, y la creación de mejores instrumentos de medición.

RESULTADOS

La tabla 1 recoge los primeros estadísticos descriptivos. Los valores t para la comparación de grupos son significativos para la confrontación más extrema, 4º-6º (t=-2,902; p=0,004). El resto de las diferencias de medias no son estadísticamente significativas 4º-5º (t=-1,483; p=0,139) y 5º-6º (t=-1,331; p=0,184).

Tabla 1. Descriptivos de las muestras, consistencia interna de las escalas y parámetros de los ítems.

	N	$\bar{X}$	S_x	α	$\bar{a}$	$\bar{b}$	$\bar{c}$
4	211	16,65	4,19	0,806	0,701	-1,03	0,107
5	186	17,30	4,56	0,828	0,954	-1,06	0,103
6	145	17,94	4,00	0,788	0,767	-1,29	0,073

La consistencia interna de las pruebas se evalúa con el alpha de Cronbach (1971). Los valores obtenidos son de 0,806 para el grupo de 4º de educación primaria, 0,828 para el grupo 5º y de 0,788 para el nivel 6º.

Evaluación de la unidimensionalidad

Para la utilización de un modelo de respuesta al ítem unidimensional es necesario comprobar empíricamente el cumplimiento de esta condición. Para ellos sometemos a un análisis de

componentes principales la matriz de correlaciones tetracóricas. La tabla 2 recoge los resultados. En ella se muestran los valores propios de los tres primeros factores y el porcentaje de varianza que explican en cada una de las muestras. Puede apreciarse que en todos los casos la varianza explicada por el primer factor supera el criterio de unidimensionalidad de Reckase (1979) según el cual éste ha de superar el 20% de la varianza total.

Tabla 2. Parámetros de los ítems obtenidos en las tres muestras.

	4º Curso			5º Curso			6º Curso		
Item	a	b	c	a	b	c	a	b	c
1	,290	-3,972	,115	,351	-3,512	,109	,325	-4,587	,080
2	,369	-3,935	,114	,371	-4,602	,107	,485	-3,694	,080
3	,486	-1,981	,115	,423	-2,038	,112	,341	-2,687	,082
4	,399	-2,161	,115	,439	-1,936	,107	,763	-2,255	,080
5	,480	-2,144	,114	,574	-1,507	,114	,466	-2,076	,078
6	,372	-1,425	,116	,442	-1,695	,116	,624	-1,453	,078
7	,590	-3,702	,112	,665	-3,241	,106	1,082	-2,964	,075
8	,564	-1,118	,107	,576	-1,309	,112	,722	-1,463	,074
9	,545	-2,940	,111	,812	-1,984	,098	,985	-2,289	,082
10	,513	-2,992	,113	2,025	-1,940	,096	,774	-3,094	,076
11	,813	-,682	,114	,527	-1,780	,102	,880	-1,268	,066
12	1,146	-1,464	,104	1,087	-1,496	,112	,886	-2,080	,073
13	,429	-,801	,114	,533	-,979	,108	,508	-1,315	,079
14	,726	-,367	,132	,676	,048	,103	,460	-1,384	,075
15	1,364	-,606	,135	1,140	-,648	,082	1,098	-1,501	,075
16	,873	-,633	,147	,788	-,961	,084	,577	-1,823	,076
17	,775	,543	,078	,944	-,090	,089	,624	-,169	,060
18	,742	,252	,116	,893	-,050	,080	,791	-,231	,058
19	,726	,777	,077	1,162	,290	,119	,750	-,039	,065
20	,810	-,469	,127	1,831	-,017	,219	1,115	-,286	,081
21	1,370	,670	,054	1,533	,363	,045	,764	,514	,068
22	,722	,388	,107	1,886	,339	,076	1,418	,082	,077
23	,794	,748	,056	1,327	,733	,110	,765	1,562	,084
24	,716	,209	,108	1,189	,696	,090	1,081	,708	,036
25	,903	1,992	,083	1,658	,724	,091	,899	1,385	,056

Estimación de los parámetros

El modelo de teoría de respuesta al ítem utilizado para la estimación de los parámetros es el modelo más general, el logístico de tres parámetros, y el procedimiento de estimación utilizado ha sido el marginal por máxima verosimilitud implementado en BILOG (Mislevy y Bock, 1990).

En ninguno de los cursos analizados aparecen ítems con valores de ajuste significativos ($p<0,01$). Los valores de los parámetros estimados en cada una de las muestras pueden observarse en la tabla 3.

Tabla 3. Estructura factorial.

	4º		5º		6º	
factores	Valor propio	%Varianza explicada	Valor propio	% Varianza explicada	Valor propio	%Varianza explicada
1	6.588	26.34%	8.498	33.92%	9.507	38.02%
2	2.641	10.56%	2.970	11,88%	3.580	14.32%
3	1.759	7.03%	2.641	10,56%	2.899	11.59%

Tabla 4. Funcionamiento diferencial de los ítems

	χ^2		
	4º-5º	4º-6º	5º-6º
1	0.1876	1.2425	0.5120
2	6.4866	0.3610	8.5050
3	4.9974	7.9437	1.4696
4	0.5867	4.4177	5.7994
5	1.2328	0.4082	1.7187
6	1.6804	3.2061	7.5318
7	0.4955	4.5371	5.4536
8	2.7978	0.4677	3.0072
9	5.9609	7.0006	1.5312
10	2.4042	1.7213	2.2532
11	21.0413**	3.8039	15.1932**
12	2.6140	5.9400	1.9331
13	0.5261	1.1638	0.2098
14	4.2921	12.1867**	15.6511**
15	1.7274	12.3383**	7.2447
16	6.4660	14.3351**	7.7208
17	4.4369	4.0154	0.9507
18	1.2306	2.3130	0.2593
19	1.2418	4.3159	3.8723
20	22.1295**	5.4728	7.2892
21	0.2128	5.9002	6.0564
22	2.7214	2.1704	1.3247
23	4.8923	21.9162**	10.1598
24	13.0590**	11.4227**	1.3945
25	10.9994	0.6016	9.03428
Total	3	5	2

Funcionamiento diferencial de los ítems

Los resultados obtenidos tras la aplicación de las técnicas de detección del funcionamiento diferencial de los ítems pueden verse en la tabla 4, donde el asterisco corresponde a un nivel de significación ($p<0,01$).

Bajo el epígrafe χ^2 aparecen los valores de este estadístico tras un proceso iterativo de purificación del criterio que converge en dos etapas. En este trabajo 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, 1993), y el cálculo de los índices de FDI con IRTDIF (Kim y Cohen, 1992).

Según estos datos en la comparación entre los cursos 4º y 5º aparecerían tres ítems con funcionamiento diferencial el 11, el 20 y el 24. En la comparación entre 5º y 6º, vuelve a aparecer el ítem 11 y además el ítem 14. Es en la comparación entre grupos más extremos, 4 y 6 donde encontramos el mayor número de ítems con funcionamiento diferencial, 5 ítems. Entre ellos, dos habían aparecido también en las comparaciones anteriores (ítem 14 e ítem 24).

Por otro lado la dirección del funcionamiento diferencial se mantiene invariable en aquellos ítems involucrados en más de una comparación. Así el ítem 11 favorece sistemáticamente a los sujetos que cursan 5º, frente a los que estudian en 4º o en 6º. Del mismo modo el ítem 24 presenta funcionamiento diferencial a favor del grupo de menor edad, sujetos de 9 años, frente al resto de sujetos en todas las comparaciones. Existe por fin un último ítem que perjudica a los sujetos de 4º y 5º, frente a los de 6º, el ítem 14. De los cuatro ítems restantes uno aparece en la comparación 4º-5º favoreciendo a los sujetos de menor edad (ítem 10). El resto de los ítems 15, 16 y 23 aparecen solo en la comparación 4º-6º, perjudicando los dos primeros a los sujetos de menor edad.

Descripción y categorización de ítems

Una vez detectados los ítems que presentan funcionamiento diferencial, el siguiente paso consistirá en la categorización de los ítems de la prueba en función de los procesos y tipos de conocimiento implicados en su resolución con el objetivo de profundizar en la búsqueda del origen del FDI y verificar nuestras hipótesis. Del examen de los contenidos se derivan tres tipologías de ítems:

1. Algorítmicos u operativos.
2. Problemas de enunciado
3. Preguntas acerca de conocimientos específicos o ítems declarativos

1. Los ítems algorítmicos son aquellos ítems en los que el contenido verbal se reduce al enunciado de la operación matemática requerida; ("suma", "resta", "divide"...). Para su resolución no es preciso poseer conocimientos de tipo lingüístico, estratégico o esquemático, sino que únicamente requieren conocimientos algorítmicos y los componentes de resolución implicados en el problema se reducen al componente de ejecución.
2. Los problemas de enunciado exigen los cuatro procesos postulados por Mayer (1982) para su resolución. Son precisos,
   - conocimientos lingüísticos y fácticos para traducir lo expresado verbalmente en una representación mental interna,
   - conocimientos esquemáticos para integrar el problema en una representación coherente de la situación descrita
   - conocimientos estratégicos para planificar un curso de acción basado en la representación
   - conocimientos algorítmicos para ejecutar la solución.

   Dentro de este grupo y atendiendo al diferente grado de especificidad de los conocimientos requeridos en el proceso de traducción del problema hay dos tipos dife-

Tabla 5. Clasificación de los ítems, FDI y curso en el que se introducen problemas análogos

Tipo		Contenido		Nº Item	FDI	Curso
Algorítmicos u operativos		**Números enteros**	+	1		4º
			-	4		4º
			x	2,3,6		4º
			:	5,8,19		4º
			Escritura nº	10		5º
		Números decimales	x	11	**5-4 5-6**	5º
			:	17		5º
Problemas de enunciado	**SIN conocimientos específcios**	**una operación**	-	13		4º
			x	7		4º
			:	9, 15	**6-4**	4º
		más de una operación	+ -	18		4º
			+ x	14	**6-4 .6-5**	4º
			x :	16	**6-4**	4º
			x x +	12		4º
	CON conocimiento específico	**una operación**	x (areas)	25		5º
			x (geometría)	22		4º
		más de una operación	: : (fracciones)	21		4º
			x : (magnitudes;)	23	**4-6**	4º
Ítems declarativos			(equiv.uds)	20	**4-5**	4º
			(nos romanos)	25	**4-5 4-6**	5º

renciados de ítems de enunciado, por un lado los problemas para los que una correcta interpretación del enunciado exige el conocimiento de ciertos contenidos matemáticos específicos como los de transformación o equivalencia de magnitudes, geometría o fracciones, y por otro los problemas de enunciado que implican operaciones aritméticas básicas, (sumas, restas, multiplicaciones, divisiones) , bien sea en uno o en varios pasos. Para esta clase de problemas no son necesarios conocimientos matemáticos adicionales.

3. Los ítems consistentes en preguntas o cuestiones. Para resolverlos se precisa de ciertos conocimientos específicos de dominio y su resolución requiere únicamente realizar el proceso de traducción del problema . En este tipo de ítems el componente procedimental es inexistente requiriendo más bien de un conocimiento exclusivamente factual, "declarativo". Anderson (1976)

A la luz de la anterior clasificación procederemos a examinar la presencia/ausencia de FDI en cada uno de los ítems así como su dirección.

Ítems algorítmicos. Con la excepción del ítem 11 hay ausencia de FDI en el bloque de ítems correspondientes a operaciones aritméticas. El ítem 11 consiste en una multiplicación de un

número decimal por 10. La introducción del algoritmo de la multiplicación de números decimales por la unidad seguida de ceros corresponde a 5º curso.

Ítems de enunciado. Los resultados siguen diferente patrón en los dos subgrupos establecidos;

1.a) Presencia de funcionamiento diferencial a favor del curso de 6º frente al curso de 4º en tres de los problemas de enunciado (ítems 14, 15 y 16). En el ítem 14 encontramos además funcionamiento diferencial del ítem (FDI) significativo en la comparación de 6º frente a 5º. Señalemos el hecho de que el ítem 14 es un problema de un único paso mientras que los ítems 15 y 16 son problemas de dos pasos. En los tres ítems existe la opción de utilizar estrategias diferentes de resolución.

1.b) Presencia de funcionamiento diferencial a favor de 5º tanto en la comparación con 4º como en la comparación con 6º.

Ítems declarativos. Aparece funcionamiento diferencial a favor del curso inferior en los dos ítems de esta categoría (ítems 20 y 24).

En la tabla 5. se detallan junto a la categoría y el contenido del ítem las comparaciones entre cursos en las cuales aparece FDI (el primer número de cada par indica el grupo favorecido sistemáticamente por el mismo), así como el curso en el que está programada la exposición de los contenidos a los que se refiere el ítem en cuestión.

DISCUSIÓN

El examen de los ítems con FDI a la luz de la anterior clasificación conduce a las siguientes apreciaciones

.- En la categoría de ítems exclusivamente algorítmicos que requieren del sujeto una operación básica y en los que la carga verbal se limita a instrucciones como multiplica o divide, no se observa FDI salvo en el caso de un ítem (número 11). Este es el único ítem que involucra números no enteros, en concreto consiste en una multiplicación de un número decimal por 100. Una resolución eficiente de este tipo de operaciones requiere la aplicación de una regla específica ("se corre la coma a la derecha tantos lugares como ceros siguen a la unidad") (ver por ej. Pereda, Pág. 54) frente al algoritmo general de la multiplicación; El aprendizaje de esta regla y en general de los algoritmos correspondientes a la multiplicación y a la división por la unidad seguida de ceros se lleva a cabo en 5º curso. En los diez ítems restantes no se presenta FDI lo que significa que la probabilidad de respuesta correcta a este tipo de ítems en sujetos con el mismo nivel de habilidad es independiente de su pertenencia a 4º, 5º o 6º curso.

.- Dentro de la categoría ítems de enunciado sin conocimientos específicos se encuentran ítems con funcionamiento diferencial y este FDI favorece sistemáticamente a los sujetos de más edad (6º curso).

Los ítems que presentan funcionamiento diferencial corresponden en todos los casos a problemas que implican las operaciones aritméticas de división (15) y/o multiplicación. (14, 16) y dos de ellos corresponden a problemas de más de un paso. Estas dos características llevan aparejadas un componente de dificultad adicional; En efecto, diferentes autores señalan el carácter considerablemente más difícil de la comprensión del significado de la multiplicación y división frente a la adición y sustracción (Brown, 1981; Dickson, Brown y Gibson, 1991; Nesher y Katriel, 1977); en cuanto al número de pasos a realizar, hay evidencia de que la utilización de dos pasos incrementa la dificultad de la tarea. (Quintero, 1983).

Los ítems de enunciado ponen en juego un amplio repertorio de capacidades y conocimientos, tales como el lingüístico, general, esquemático, estratégico y algorítmico, cuya integración es necesaria para su resolución correcta. Las diferencias interpersonales e intergrupales, en nuestro caso entre los cursos 4º, 5º y 6º, pueden darse en todos y en cada uno de los procesos exigidos.

Así, en el nivel de la traducción los sujetos pueden diferir en la capacidad para comprender las expresiones lingüísticas, en la fase de integración en su conocimiento de los distintos tipos de problemas que se presentan mediante palabras, en la fase de planificación pueden diferenciarse en las estrategias generales de resolución de problemas, y por fin en la fase de ejecución, pueden diferir en el grado de sofisticación, corrección y automaticidad en el cálculo de los algoritmos de operaciones básicas (Mayer, 1982). La mejora con la edad de la habilidad para la resolución de problemas de enunciado viene acompañada por 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), lo cual significa que el conocimiento lingüístico se incrementa con la edad (Greeno, 1980). 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.

Los resultados muestran que cuando se observa funcionamiento diferencial éste atañe en mayor número de ocasiones a las comparaciones entre los cursos extremos del rango muestral (4º y 6º), es decir, afecta a los sujetos que mayores divergencias observarán en las capacidades o habilidades mencionadas.

En el subgrupo de ítems de enunciado que demandan conocimientos específicos se observa funcionamiento diferencial únicamente en el ítem 23. Este ítem involucra dos operaciones y exige además el conocimiento de la correspondencia entre unidades de tiempo, en concreto entre horas y minutos. Es remarcable el hecho de que el FDI favorezca al curso 4º frente a 6º, en contraste con la dirección del FDI para el grupo de ítems anterior. Siendo el conocimiento específico requerido por el ítem reforzado a lo largo de todo el proceso educativo tal vez sea la proximidad de la exposición de contenidos lo que explique esta circunstancia.

.- En cuanto a los ítems declarativos se observa FDI tanto en el caso del ítem 24 como en el 20. En el ítem 24 aparece siempre a favor del 4º curso frente a los otros dos (5º y 6º). El examen de la distribución de contenidos a lo largo de los cursos permite constatar que es precisamente en 4º curso en el que se instruye a los niños en la escritura de números romanos y tras la revisión de textos teóricos y de problemas, se comprueba la ausencia de referencia a dichos contenidos en cursos superiores. Por lo tanto 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 la superioridad sistemática de los sujetos de curso inferior. En el caso del ítem 20 también corresponde a 4º curso la exposición de contenidos relativos a la equivalencia entre unidades de longitud, aunque en este caso no parece plausible sostener la ausencia de refuerzo en el conocimiento.

En conclusión diríamos que existen fuentes de sesgo a tener en cuenta cuando una misma prueba es aplicada a sujetos de diferentes cursos, y que éstas parecen estar relacionadas con la naturaleza y contenido de los ítems. Si bien existe un tipo de problemas complejos, que requieren una realización experta (Kintsch y Greeno, 1985), y éstos favorecen a los sujetos de más edad, en el caso de aquellos ítems que exigen conocimientos fácticos como relaciones, definiciones, fórmulas, será preciso tener en cuenta la influencia que sobre la retención ejerza el tiempo transcurrido entre la instrucción y la ejecución de la prueba.

Digamos además que existe cierta evidencia de que los conocimientos procedimentales son retenidos mejor a largo plazo que los conocimientos declarativos (Conway, Cohen y Stanhope,1991) . Si los conocimientos declarativos son más susceptibles de olvido, entonces se verán beneficiados en mayor medida por la proximidad de la exposición , y esta aseveración es consistente con los resultados obtenidos en nuestro estudio que evidencian la influencia de la secuenciación de contenidos en aquellos ítems que exigen conocimientos declarativos junto a la ausencia de FDI en los ítems de contenido preferentemente procedimental o algorímico.

Si bien somos conscientes del carácter exploratorio de nuestro análisis y de la conveniencia del diseño de estudios ad-hoc para obtener evidencias más concluyentes, confiamos en que el

presente trabajo suponga una llamada de atención acerca de cuestiones que raramente son tenidas en cuenta por los constructores de test. La adopción de la metodología derivada del estudio del funcionamiento diferencial para la detección del sesgo contribuiría a garantizar la equidad del proceso de medida, permitiendo descartar aquellos ítems contaminados por variables espurias y seleccionando aquellos que, siendo útiles para la medición del constructo, preserven la imparcialidad de contenidos mencionada por Title (1982).

Por otro lado nos parece conveniente la adopción de un enfoque analítico de las pruebas psicométricas, centrado en el contenido y en la naturaleza de los ítems y en los procesos implicados, más allá del mero cálculo de índices psicométricos. La importancia de un planteamiento integrador de la teoría cognitiva y las técnicas de medida es destacado por Glaser (1998) cuando afirma que "el valor de los sistemas de evaluación que establecemos depende íntimamente de nuestro conocimiento de cómo los humanos aprenden y adquieren conocimientos y habilidades." (pp. 36) y desde esta perspectiva hemos planteado nuestro trabajo.

Referencias

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

Ausubel, D.P., Novak, J.D: y Hanesian H. . (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)

Brown, M. (1981) . Number Operations, en K. Hart (ed.) *Children's understanding of mathematics.* John Murray,

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

Bontempo, R.(1993). Translation fidelity of psychological scales: An item response theory analysis of an individualism-collectivism scale. *Journal of cross-cultural psychology, 24* (2), 149-167.

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.

Conway, M.A.; Cohen, G. y Stanhope, N. On the very long term retention of knowledge acquired trough formal education: twelve years of cognitive psychology. *Journal of Experimental Psychology: General, 120,* 395-409

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

Drasgow, F.(1987). Study of the Measurement Bias of two Standarized Psychological Tests. *Journal of Applied Psychology, 72*(1), 19-29.

Ellis, BB., Becker.P. y Kimmel, H.D.(1993). An item response theory evaluation on an english version of the Trier Personality Inventory (TPI). *Journal of Cross-cultural psychology,* 24(2), 133-148.

Ellis, B.B. y Kimmel, H.D.(1992). Identification of unique cultural response patterns by means of Item Response Theory. *Journal of Applied Psychology, 77*(2), 177-184.

Elosúa, P. y López.A(1999) Funcionamiento diferencial de los ítems y sesgo en la adaptación de dos pruebas verbales. *Psicológica* 20(1),23-40

Elosúa, P. López, A y Egaña, J. (en prensa) Idioma de aplicación y rendimiento en una prueba de comprensión verbal.

Glaser, R. (1998). Pericia y evaluación. En Wittrock, M.C. y BAker, E. L. (comp.) Test y cognición. Investigación cognitiva y mejora de las pruebas psicológicas. Barcelona. Paidós, 1998 (Orig. Testing and cognition. N. Jersey. Prentice-Hall, 1991)

Gomez, J. y Navas, M.J.(1998) Impacto y funcionamiento diferencial de los ítems respecto al genero en una prueba de aptitud numerica. *Psicothema, 10*(3) 685-696

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. y Swaminathan, H.(1985). *Item response theory: Principles and Applications.* Boston: Kluwer-Nijhoff.

Hulin, C.L. y Mayer, L. (1986) Psychometric equivalence of a translation of the job descriptive index into hebrew. *Journal of applied psychology, 71*(1), 83-94.

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.

Kintsch, W. y Greeno, J.G:, (1985) Understanding and solving word arithmetic problems, *Psychological Bulletin*, 92 (1), 109-129

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

Linn, R.L., Levine, M.V., Hastings, C.N. y Wardrop, J.L.(1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*(2), 159-173.

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. (1982). Mathematical ability. 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)

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.

Muñiz, J.(1997). *Introducción a la teoría de respuesta a los ítems.* Madrid: Pirámide.

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

Nesher, P. y T. Katriel (1977) A Semantic Analysis of Addition and Subtraction Word Problems in Arithmetic, *Educational Studies in Mathematics, 8*, 251-269

Padilla, J.L., Pérez, C. y González, A. (1999) Diferencias instruccionales y funcionamiento dierencial de los ítems: Acuerdo entre el método Mantel-Haenszel y la regresión logística. *Psicológica, 19*(1), 201-215.

Pereda, L. (1993) Matemáticas 5. n, Magnitudes, Estadística y Geometría. San Sebastian Erein.

Quintero, -A.H. (1983) The Role of Conceptual Understanding in Solving Word Problems: Two-Step Problems. *Congreso de la Educational Research Association* (Montreal, Canada, Abril, 1983).

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.

Reckase, M.D.(1979): Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics, 4*, 207-230.

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.

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.

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

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.

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.