Psicológica (1999) 20, 121-134.

Análisis de la dificultad de un test de matemáticas mediante un modelo componencial[1]

Eulogio Real (**), Julio Olea (*), Vicente Ponsoda (*), Javier Revuelta (*) y Francisco J. Abad (*)

(**) Universidad de Santiago de Compostela

(*) Universidad Autónoma de Madrid

En este trabajo se delimitan los componentes de dificultad que intervienen en la resolución de los ítems de un test de matemáticas que incluye 66 ítems de operaciones con números enteros positivos y negativos. Se estudia el ajuste al modelo de Rasch y se estiman los parámetros de dificultad de los componentes mediante el modelo LLTM de Fischer. Se retuvieron 8 de los 10 componentes propuestos. Los parámetros de dificultad predichos por el modelo LLTM y los estimados mediante el modelo de Rasch obtuvieron una relación lineal positiva elevada (r=0.8783). Los resultados de este estudio preliminar animan a seguir trabajando en el desarrollo de un sistema de Generación Automática de Items (GAI).

Palabras clave: modelo de Rasch, modelos componenciales, test de matemáticas, generación automática de ítems.

Desde hace no mucho tiempo se están produciendo serios intentos de acercamiento entre la Psicología Cognitiva y la Psicometría. Además de tener en cuenta el resultado en un ítem (acierto o fallo) para estimar el nivel de rasgo de una persona, han surgido modelos psicométricos que pretenden incorporar los diferentes componentes o procesos cognitivos implicados en su resolución. Un claro ejemplo de este nuevo enfoque lo constituyen los denominados "modelos componenciales" (Prieto y Delgado, 1999; Van der Linden y Hambleton, 1997). Básicamente, un modelo componencial requiere: a) un análisis de las operaciones mentales (componentes cognitivos) que intervienen en la resolución de los ítems, y b) un modelo matemático que estime la probabilidad de acertar un ítem teniendo en cuenta sus propiedades estructurales y el nivel de conocimiento del sujeto. Determinar las propiedades estructurales de los ítems significa por tanto delimitar el tipo, la cantidad y el orden de los procesos que intervienen en su resolución. El modelo psicométrico sirve para estimar el grado en que los diferentes componentes contribuyen a la dificultad del ítem.

Uno de los modelos componenciales más utilizados es el modelo logístico lineal de rasgo latente (Linear Logistic Latent Trait Model, LLTM), de Fischer (1973, 1997), según el cual la dificultad final del ítem es resultado de la suma de las dificultades de los componentes implicados. El modelo

[1] Agradecimientos: Este trabajo ha sido financiado por los proyectos DGICYT PS94-0040 y DGES PB97-0049. Dirección de contacto: Julio Olea. Facultad de Psicología. Universidad Autónoma de Madrid. 28049-Madrid. E-Mail: Julio.Olea@uam.es

LLTM descompone, mediante una combinación lineal, la dificultad *b* del modelo de Rasch en una serie de *j* componentes de dificultad η que se ejecutan *f* veces, más una constante *c* de escalamiento. Formalmente:

$$b = \sum_{i=1}^{j} f_i \eta_i + c$$

Si se delimitan de forma correcta los componentes, cabe esperar una relación lineal elevada entre los parámetros *b* estimados por el modelo de Rasch y los parámetros *b** predichos por el modelo LLTM. Cuando esto es así, las posibles aplicaciones de los resultados obtenidos son muy variadas: resulta posible, por ejemplo, construir tests con demandas cognitivas conocidas, diseñar un test adaptativo informatizado (TAI) que incorpore la información sobre los diferentes componentes de dificultad o, lo que parece muy interesante, crear ítems con propiedades psicométricas conocidas sin necesidad de someterlos a un proceso empírico de calibración. Esta última aplicación se conoce como Generación Automática de Items (GAI). La GAI consiste en crear ítems automáticamente mediante determinados algoritmos, lo cual puede tener indudables ventajas: a) a nivel teórico, proporciona un sustrato cognitivo a la capacidad evaluada por el test, mediante un modelo que identifica los componentes esenciales de dicha capacidad, así como la importancia relativa de cada uno, b) al centrarse en los procesos de resolución de las respuestas, la GAI muestra un mayor interés por la validez de los tests (Bejar, 1993; Embretson, 1995; cf. Revuelta y Ponsoda, 1998), y c) al no necesitar de un gran banco de ítems que elaborar y calibrar, la GAI puede representar a nivel aplicado una gran reducción de costes. El tipo de tareas al que se han aplicado estas estrategias es ya bastante amplio (cf. Irvine, Dann y Anderson, 1990; Revuelta y Ponsoda, 1998, 1999), y abarcan, entre otros, aspectos tales como la lectura de números romanos (Solano-Flores, 1993), la resolución de problemas espaciales con figuras tridimensionales (Bejar, 1986), la evaluación de las destrezas en lenguaje escrito (Bejar, 1988, 1996), la resolución de problemas de matemáticas (Medina-Díaz, 1993; Meisner, Luecht y Reckase, 1993), la resolución de matrices tipo Raven (Hornke y Habon, 1986) y la resolución problemas de análisis lógico (Revuelta y Ponsoda, 1998).

En la presente investigación se estudia el ajuste del modelo LLTM a un test de matemáticas. La idea es delimitar los diferentes componentes de dificultad que intervienen en la resolución de los ítems y comprobar el ajuste al modelo, como primeras fases para establecer un procedimiento de GAI de matemáticas.

MÉTODO

Material. El test, llamado "prueba de signos" (Alonso Tapia y Olea, 1997) forma parte de una batería más amplia de evaluación de conocimientos de matemáticas para alumnos de primer ciclo de la ESO. Contiene 66 ítems de similar formato: operaciones del tipo *a* * *b* = *c*, donde *a*, *b*, y *c* son números enteros positivos o negativos, y "*" uno de los cuatro operadores aritméticos posibles (suma, resta, multiplicación y división). La respuesta del sujeto consiste en identificar la corrección o incorrección del resultado *c* que se proporciona. En el contexto de la evaluación de conocimientos

matemáticos, este formato de preguntas se conoce como "sentencias canónicas de verificación" (Maza, 1989). En la prueba de signos, cuando un resultado *c* es incorrecto, su valor no se establece aleatoriamente, sino que se consideran los errores más frecuentes que se cometen con este tipo de operaciones (Dickson, Brown y Gibson, 1991). 24 ítems son sumas, 24 restas, 9 productos y 9 divisiones. Las 4 combinaciones posibles entre los dos signos de los términos *a* y *b* (++, --, +- y -+) se dan en todas las operaciones.

Muestra. La prueba de signos se aplicó a una muestra de 221 alumnos de 7º de EGB y 1º de la ESO, de tres colegios concertados de Madrid.

Delimitación de componentes. Al analizar las propiedades estructurales que se incluyen en los ítems de la prueba de signos, se encontraron al menos dos características a resaltar: a) no incorporan otro tipo de variaciones en el formato de las preguntas (por ejemplo el orden de los términos, la cantidad de operaciones a realizar o el lugar donde se ubica la incógnita) que también pueden intervenir en su dificultad, y b) la prueba no incluye algunas de las propiedades estructurales de las operaciones aritméticas involucradas (por ejemplo la conmutatividad en la suma y el producto o la asociatividad y distributividad entre algunas de ellas). Parte de estas limitaciones tienen que ver con el tipo de población para la que fue inicialmente ideada (alumnos de primer ciclo de la ESO).

En relación a la delimitación de componentes, algunos teóricos del aprendizaje de las matemáticas (p. ej. Fuson, 1992; Schwarz, Kohn y Resnick, 1992) reconocen lagunas importantes en el estudio de los procesos implicados en la resolución de problemas con números enteros negativos, y más concretamente en los algoritmos de cálculo. Sin embargo, sí han sido objeto de estudio los procesos implicados en las diferentes operaciones cuando se aplican a números enteros positivos (p. ej. English y Halford, 1995; Grows, 1992; Leder, 1992 o Sloboda y Rogers, 1987).

El planteamiento que se asume en este estudio tiene la base en estos trabajos y en la teoría sobre esquemas. Según Marshall (1993) un esquema incluye cuatro tipos de conocimientos: (1) conocimiento de los rasgos distintivos de un fenómeno o situación, (2) conocimiento de las condiciones que los delimitan y, por tanto, de la posibilidad de aplicar determinados procedimientos para resolver problemas, (3) conocimientos relativos a la planificación de la solución de un problema de una categoría dada (estrategias de resolución del problema), y (4) conocimientos relativos a los algoritmos y reglas de cálculo necesarias para ejecutar el proceso de solución.

Así pues, si se organiza el contenido de los problemas de la prueba de signos atendiendo a estos cuatro tipos de conocimiento, se obtienen los elementos u operaciones mentales implicados en cada problema. El primer tipo de conocimiento establece distinciones entre los cuatro tipos de operaciones aritméticas (suma, resta, multiplicación y división). Es de suponer que la dificultad es diferente para cada tipo de operación. El segundo tipo de conocimiento establece distinciones entre varias situaciones posibles

en que puede plantearse el problema: (a) que el término *a* sea positivo o negativo; (b) que el término *b* sea positivo o negativo; (c) que el tamaño del término *a* sea mayor o menor que el tamaño del término *b*, en valor absoluto. El tercer tipo de conocimiento establece una serie de estrategias que permiten llegar a la solución, dependiendo del modo en que se dispongan los elementos del problema en función de los condicionantes anteriores (cambiar el signo de los términos *a*, *b*, ó *c*, cambiar el signo del operador, intercambiar entre sí los términos *a* y *b*). El cuarto tipo de conocimiento se refiere a los conocimientos que los alumnos deben tener para llevar a cabo una operación aritmética simple, previa a la obtención de la solución.

El primer paso, por tanto, en el estudio del proceso de resolución de los ítems consistió en la determinación de las posibles combinaciones de todos los elementos de cada problema. Los factores implicados en la aparición de estas combinaciones fueron cinco para las operaciones de suma y resta, y cuatro (uno menos) para las operaciones de multiplicación y división. Para las operaciones de suma y resta, los cinco factores fueron: (1) tamaño relativo de *a* y *b*, tomados éstos en valor absoluto ($|a|>|b|$ ó $|b|>|a|$); (2) signo de *a* (positivo o negativo); (3) signo de *b* (positivo o negativo); (4) operador implicado (suma o resta); y (5) cambio del signo del resultado. En el caso de la multiplicación y división, el factor referido al tamaño relativo de *a* y *b* no era relevante (no influye en el proceso de multiplicar y, para la división, *a* siempre es mayor que *b* en valor absoluto). En cuanto al cuarto factor, aquí los operadores implicados eran la multiplicación y la división.

El proceso de resolución de los problemas se planteó como un árbol de decisión no iterativo. Inicialmente, se plantearon 12 posibles componentes o procesos en el desarrollo de la resolución de un problema; los cuatro primeros se refieren a las diferencias entre los cuatro tipos de operaciones aritméticas; los tres siguientes se refieren a las condiciones en que se plantea la operación, mientras que los cinco siguientes se aplican a la resolución en sí del problema. Los procesos planteados fueron los siguientes:

- *suma* - Operador suma.
- *resta* - Operador resta.
- *multipl.* - Operador multiplicación.
- *división* - Operador división.
- $a \geq 0?$ - Comprobación de que *a* es positivo.
- $b \geq 0?$ - Comprobación de que *b* es positivo.
- $|a| \geq |b|?$ - Comprobación de que *a* es mayor o igual que *b* en valor absoluto.
- $\pm a$ - Cambia el signo de *a*.
- $\pm b$ - Cambia el signo de *b*.
- $\pm *$ - Cambia el operador suma por resta, o viceversa.
- $a \leftrightarrow b$ - Intercambia de posición los miembros *a* y *b*.

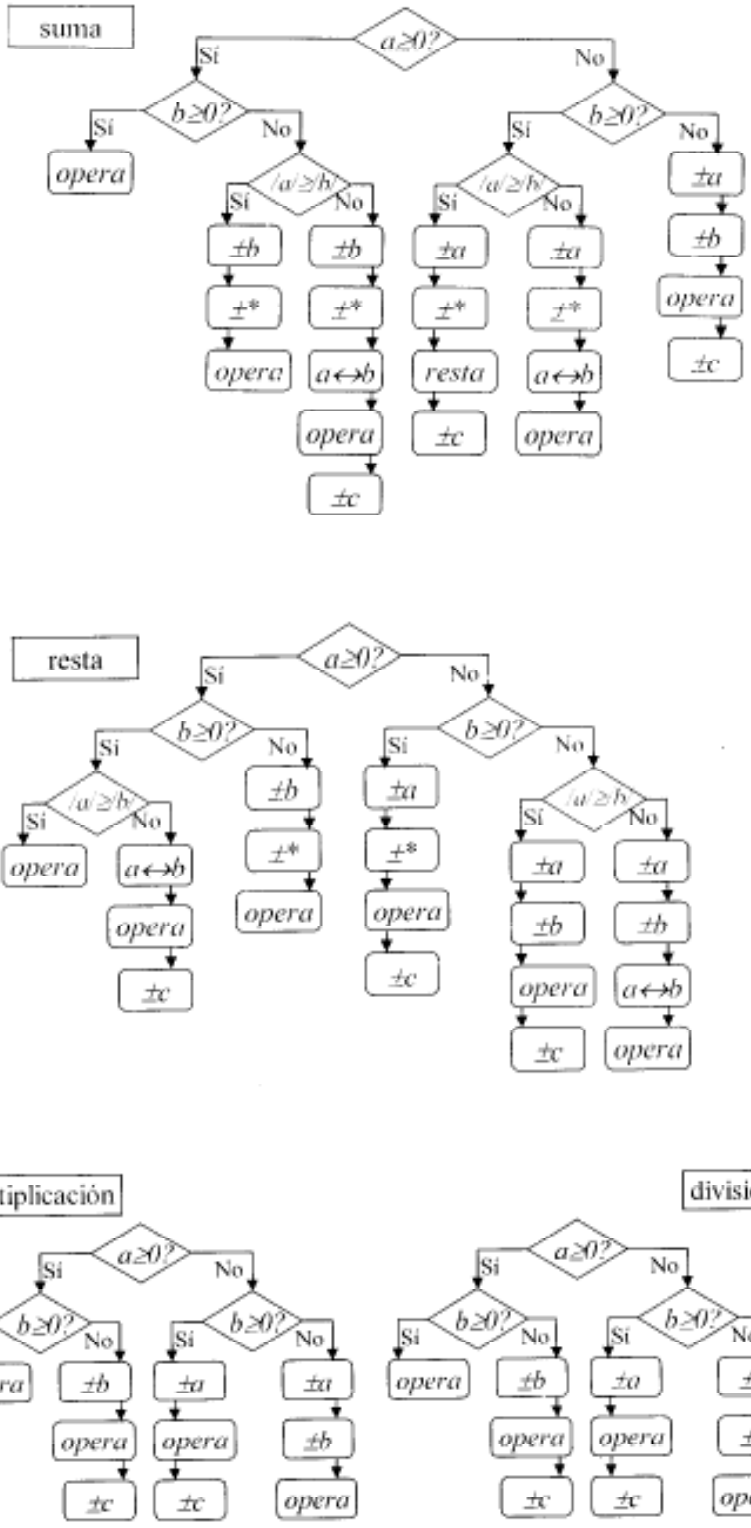
· $\pm c$ - Cambia el signo del resultado.

En la página siguiente se muestran los árboles de decisión para cada uno de los operadores implicados.

Establecidos de esta forma los árboles de decisión, dos de los componentes ($a \geq 0?$ *y* $b \geq 0?$) se desestimaron porque resultaban necesarios para la resolución de cualquier tipo de ítem. Además, este tipo de condiciones resultaban redundantes con otros procesos ya contemplados ($\pm a$ y $\pm b$, respectivamente). De este modo, cada uno de los ítems de la prueba de signos quedó asociado a un vector binario de 10 dígitos, donde cada uno corresponde a un componente de dificultad distinto. En el caso de los cuatro primeros (*suma, resta, multiplicación* y *división*), el 1 indica que la operación correspondiente es la involucrada; las otras tres operaciones recibirán un 0. Para los seis procesos restantes ($|a| \geq |b|$, $\pm a$, $\pm b$, $\pm *$, $a \leftrightarrow b$ y $\pm c$), el valor 1 indica que se requiere ese proceso en la resolución del problema, mientras que el 0 indica que el proceso no es necesario.

Estudio psicométrico del test. Desde los desarrollos de la Teoría Clásica de los Tests, se obtuvo la dificultad de los ítems, su discriminación (correlación biserial-puntual con el total de la prueba) y la consistencia interna del test. Adicionalmente, se estudió el grado de unidimensionalidad obteniendo los componentes principales a partir de la matriz de correlaciones tetracóricas entre los ítems.

Aunque el tamaño muestral no es óptimo, dada la robustez del modelo a esta circunstancia (Barnes y Wise, 1991) se procedió a comprobar el ajuste de los ítems al modelo de Rasch y la precisión global de la prueba, mediante el programa RASCAL (Assessment Systems Corporation, 1989). Adicionalmente, y siguiendo las recomendaciones de Hambleton, Swaminathan y Rogers (1991), se realizaron análisis para comprobar: a) la invarianza de las estimaciones de θ (correlación entre las estimaciones con los ítems pares e impares, así como con los ítems fáciles y difíciles),



b) la invarianza de los parámetros *b* (correlación entre los valores estimados con dos grupos extremos de habilidad), y c) la suposición de parámetros de pseudoazar iguales a cero (estudio de las tasas de acierto en los ítems difíciles para los sujetos de menor habilidad).

Una vez realizadas estas comprobaciones, se estimaron los parámetros de dificultad de los componentes, mediante el programa LPCM-Win (Fischer y Ponocny-Seliger, 1997). De este modo se aplicó el modelo LLTM a los datos correspondientes a los 221 sujetos que respondieron los 66 ítems de la prueba de signos. Como matriz de datos se introdujo la matriz binaria (221x66) de aciertos y errores de los sujetos en los 66 ítems de la prueba. Como matriz de parámetros se introdujo la matriz binaria (66x10) de procesos para cada ítem.

Para medir el ajuste del modelo LLTM, se obtuvieron 3 indicadores diferentes: a) la correlación entre los parámetros de dificultad estimados con RASCAL y los predichos por el modelo LLTM, b) la raíz cuadrada de las diferencias cuadráticas medias entre ambas variables (RMSD), y c) el estadístico χ^2 de Andersen (Fischer, 1997), que se obtiene a partir de la razón de verosimilitudes condicionadas entre los parámetros *b* estimados por los dos modelos.

RESULTADOS

Aplicación del modelo de Rasch y comprobaciones adicionales. La prueba de signos resultó asequible para la muestra seleccionada: 18 ítems obtuvieron una proporción de aciertos entre 0.333 y 0.666, mientras que 48 obtuvieron una proporción mayor. Entre los ítems de mayor dificultad se encontraban restas con al menos uno de los términos negativos. Todas las correlaciones ítem-total fueron positivas (media=0.358, σ=0.15); 41 de los 66 ítems obtuvieron una correlación con el total dentro del rango $\pm 1\sigma$. El coeficiente α del test fue de .91. El autovalor asociado al primer componente principal explicó un 26 % de la varianza total.

En la figura 1 se incluye la distribución de frecuencias de los parámetros de dificultad y de sus errores típicos de estimación. Los errores típicos de los parámetros *b* estimados obtuvieron valores entre .184 y .386, excepto para los cinco ítems con menor dificultad (entre .976 y .501). Para 44 de los 66 ítems se obtuvieron valores de χ^2 no significativos (α=.01). La dificultad media de los ítems fue de 0, con una desviación típica de 1.59. La media para las estimaciones de θ fue 2.11 y la desviación típica 1.42.

Al comparar las estimaciones de habilidad realizadas para ítems pares e impares, la correlación obtenida fue de .847 (p<.01). Sin embargo, al comparar las estimaciones realizadas para ítems fáciles y difíciles, la correlación obtenida entre estimaciones fue .303 (p<.01). La representación gráfica de la relación entre las estimaciones muestra una sobreestimación de los valores de θ cuando se utilizan ítems fáciles, de modo que aparecen gran cantidad de sujetos con estimaciones de habilidad elevadas cuando se usan ítems fáciles, y cuya habilidad estimada desciende a niveles medios, e incluso bajos, cuando se usan ítems difíciles.

La segunda comprobación de invarianza consistió en comparar el valor de *b* para cada uno de los ítems obtenido para el grupo de sujetos con baja

habilidad estimada (por debajo de la media) y para el grupo con alta habilidad estimada. La correlación entre los valores fue de .908 (p<.01).

La última comprobación, encaminada a estudiar la existencia de adivinación en los ítems, consistió en examinar el porcentaje de aciertos por parte de los 132 sujetos con estimaciones de habilidad por debajo de la media (θ<2.11) en los 9 ítems más difíciles del test (aquellos con un valor de b mayor de 2). Los porcentajes de acierto se compararon con los estimados teóricamente por la distribución binomial bajo el supuesto de acierto aleatorio. El 43.9% de los sujetos no acertó ninguno de los ítems (b(0,9,.5) = 0.2%), el 18.9% acertó sólo uno de ellos (b(1,9,.5) = 1.76%), el 15.2% sólo dos de ellos (b(2,9,.5) = 7.03%), el 9.1% acertó tres ítems (b(3,9,.5) = 16.41%) y el 4.5% acertó cuatro ítems (b(4,9,.5) = 24.61%). En total, el 91.7% de los sujetos acertaron cuatro o menos de cuatro ítems de los nueve más difíciles, frente al 50% que sería esperable al azar.

Aplicación del modelo LLTM. La correlación entre los valores b estimados por el modelo LLTM a partir de los componentes seleccionados y los valores b determinados por el modelo de Rasch fue de .8783, lo que indica que las estimaciones de ambos modelos mantienen un orden bastante parecido. El valor χ^2 de Andersen fue 808.29 (g.l.=55, p<0.05), lo que indica que los parámetros b estimados mediante el modelo LLTM difieren significativamente de los predichos por el modelo de Rasch. El valor de RMSD resultó ser 0.78 (0.48 para las b tipificadas), lo que indica cierto grado de desemejanza entre las cuantías estimadas para los parámetros.

Las estimaciones de los parámetros resultaron significativas (p<0.01) para 8 de los 10 componentes considerados. Los errores típicos de estimación oscilaron entre 0.0578 y 0.0993, lo que indica una buena precisión de las estimaciones. Los componentes o procesos que no alcanzaron la significación fueron $\pm a$ y el operador *división*. En la tabla 1 se muestran los valores de los parámetros estimados junto con sus errores típicos, puntuaciones z y significación.

FIGURA 1A: DISTRIBUCIÓN DEL PARÁMETRO DE DIFICULTAD ESTIMADO EN EL MODELO DE RASCH

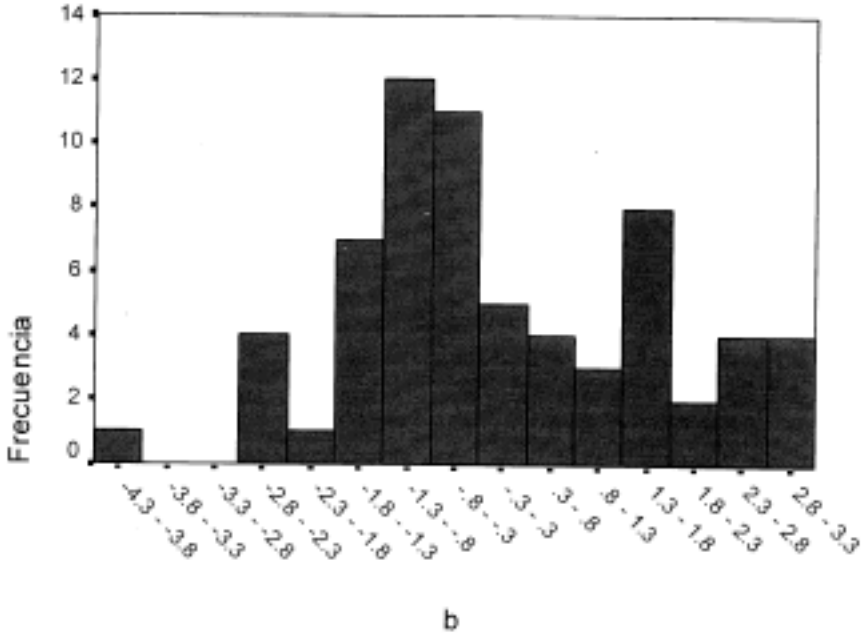


FIGURA 1B: DISTRIBUCIÓN DE LOS ERRORES TÍPICOS DE ESTIMACIÓN DE b

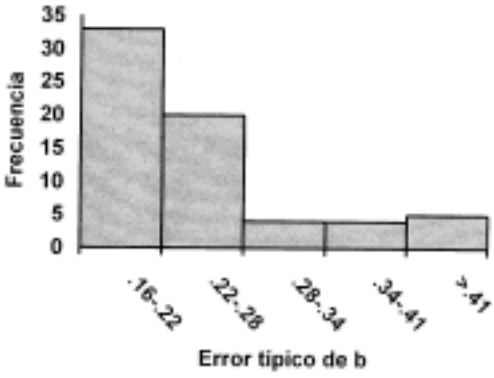


Tabla 1: Estimación de parámetros en el modelo LLTM.

Proceso	**Parámetro**	**Error típico**	**Z**	**Significación**
|a|≥|b|?	-0.5332	0.0693	7.6903	P<.01
±a	0.0883	0.0805	1.0972	n.s.
±b	0.9979	0.0578	17.2761	P<.01
±*	0.4377	0.0602	7.2695	P<.01

a ↔ b	-0.2246	0.0993	2.2613	P<.05
±c	0.5031	0.0790	6.3682	P<.01
Suma	-0.8180	0.0774	10.5639	P<.01
Resta	1.5356	0.0685	22.4211	P<.01
Multipl.	-0.7404	0.0907	8.1612	P<.01
División	0.0229	0.0835	0.2739	n.s.

Los resultados parecen indicar, pues, un buen ajuste del modelo planteado, dado que casi todos los componentes contribuyen a la dificultad de los ítems. El principal predictor de la dificultad de un ítem es la resta, de modo que cuando éste es el operador involucrado, la dificultad del ítem es mayor. El siguiente mejor predictor de la dificultad de un ítem es $\pm b$. Los siguientes mejores predictores son las operaciones suma y multiplicación; cuando estos son los operadores involucrados, la dificultad del ítem es menor. De los siguientes predictores, de menor importancia, algunos corresponden a condiciones que, si se dan, disminuyen la dificultad del ítem ($|a| \geq |b|$? y $a \leftrightarrow b$), mientras que otros se refieren a operaciones necesarias con ítems más difíciles ($\pm$* y $\pm c$).

DISCUSIÓN

Este trabajo representa el primer paso para el desarrollo de un sistema GAI para la evaluación del rendimiento en operaciones con números enteros. El resultado más sobresaliente es el elevado nivel de ajuste lineal obtenido entre los modelos de Rasch y LLTM (por encima del 77% de varianza común), lo cual parece indicar que los diferentes componentes delimitados resultan exhaustivos para predecir la dificultad de cuestiones similares a las de la prueba de signos. En comparación con trabajos previos realizados sobre tests diferentes (Hornke y Habon, 1986; Bejar, 1993, Embretson, 1993; Revuelta y Ponsoda, 1998) el ajuste lineal obtenido es sin duda el superior, y nos hace albergar esperanzas de que un sistema GAI pueda funcionar razonablemente bien. Respecto a los otros dos indicadores de ajuste, el mismo Fischer (1997) reconoce la dificultad de obtener valores χ^2 no significativos al calcular la razón entre verosimilitudes condicionadas, algo que resulta evidente incluso en trabajos donde de forma inequívoca se cumplen los supuestos del modelo de Rasch (Medina-Díaz, 1993). Los valores RSMD obtenidos informan también de una cierta discrepancia entre los parámetros de dificultad que estiman ambos modelos.

La alta correlación obtenida entre los parámetros de dificultad, cuando se estiman en las muestras de habilidad baja y alta, representa un indicador empírico del elevado grado en que se cumple la propiedad de invarianza de la TRI en la estimación de parámetros. Sin embargo, determinados resultados del estudio psicométrico de la prueba de signos, y más concretamente los relativos al ajuste que manifiesta al modelo de Rasch, aconsejan tomar ciertas precauciones antes de desarrollar los algoritmos de creación de ítems según una estrategia GAI. En primer lugar, aunque la varianza explicada por el primer autovalor supera alguno de los criterios mínimos de

unidimensionalidad propuestos (p.ej. Reckase, 1979), no puede considerarse un resultado completamente satisfactorio (en Cuesta, 1996, se informa de otros criterios más restrictivos) lo que puede incidir en las estimaciones de los parámetros de dificultad y de habilidad. En segundo lugar, no se puede desdeñar el pobre ajuste de algunos ítems al modelo de Rasch. En este momento, se ha desestimado su eliminación de la prueba, por dos razones fundamentales: a) porque, como es sabido, el criterio estadístico suele ser una prueba muy exigente como criterio de selección de ítems, y b) porque el objetivo de la investigación no es tanto garantizar la bondad psicométrica del test (aspecto exigible para su aplicación real en contextos educativos) como estudiar los procesos intervinientes en su resolución. En relación con este último argumento, puede preverse incluso un mejor ajuste entre el modelo de Rasch y el LLTM si se consiguiera un mejor grado de ajuste de los ítems al modelo logístico de un parámetro. Resulta evidente, por el valor de la media de habilidad estimada y por la tasa de aciertos obtenida en los ítems, que la prueba de signos resulta muy accesible para la muestra seleccionada. Ya se ha comentado que algunos ítems, los más desajustados, resultan extremadamente fáciles para la muestra seleccionada. Este dato no es independiente del pobre indicador de invarianza obtenido cuando se comparan las estimaciones con ítems fáciles y difíciles.

De cualquier forma, cabe destacar el bajo nivel de acierto aleatorio obtenido en este trabajo, algo que en principio hubiera resultado difícil de aventurar en ítems con dos opciones de respuesta. Los resultados obtenidos por el grupo de baja habilidad en los ítems más difíciles representan una prueba del escaso acierto aleatorio que se ha producido en la prueba de signos, requisito necesario para la aplicación del modelo de Rasch. Una posible explicación es la edad de los sujetos (entre 12 y 14 años), en la que quizás no se asimila los posibles beneficios de las respuestas aleatorias en pruebas de opción múltiple. Una segunda posible explicación tiene que ver con la lógica mediante la que se diseñaron los ítems: cuando como resultado se daba una cantidad incorrecta, siempre podía llegarse a ella cometiendo algún error en el proceso de resolución. Por ello, cabe pensar que los sujetos de menor nivel, lejos de emitir respuestas aleatorias, normalmente daban como bueno un resultado que coincidía con el que ellos obtenían de forma equivocada.

De cara a trabajos futuros, deberían elaborarse ítems más ajustados, que cumplieran de forma más clara los requisitos de la TRI y del modelo de Rasch en particular, y que incorporasen componentes adicionales no considerados en la prueba de signos, algo que por otra parte contribuirá a incrementar la dificultad del test. Respecto al formato de los ítems, la manera de evitar completamente los problemas de acierto aleatorio (que pueden poner en entredicho el supuesto de parámetro $c=0$) es establecer un formato de respuesta construida con corrección automática. En cuanto a las comprobaciones de ajuste al modelo LLTM, sería deseable incorporar el procedimiento "Quadratic Assignment" (utilizado, por ejemplo, en el trabajo ya citado de Medina-Díaz), para determinar el grado en que las respuestas empíricas de los sujetos se corresponden con el modelo de procesamiento propuesto. Finalmente, debe subrayarse que el tamaño muestral del presente

estudio se encuentra ligeramente por debajo de los requisitos mínimos establecidos por ciertos especialistas (Barnes y Wise, 1991). En trabajos futuros debería considerarse el incremento de la ratio sujetos/ítems, con objeto de intentar reducir los errores de estimación de los parámetros de dificultad.

CONCLUSIONES

A partir de los resultados obtenidos, las principales conclusiones que se pueden extraer son:

a) Respecto a la prueba de signos, que convendría incorporar ítems de mayor dificultad con las propiedades estructurales de las distintas operaciones aritméticas.

b) Respecto a los componentes de dificultad propuestos, que han resultado muy exhaustivos para predecir la dificultad de los ítems. Convendría sin embargo incorporar otros componentes adicionales.

c) Respecto al modelo de Rasch, que ítems de 2 opciones de respuesta pueden tener tasas de adivinación pequeñas en determinadas poblaciones y en cierto tipo de ítems, si se piensa bien la redacción de las opciones incorrectas.

d) Respecto a las medidas de ajuste entre el modelo de Rasch y el LLTM, que se puede obtener una relación lineal elevada entre los parámetros de dificultad estimados mediante ambos modelos, aunque el ajuste al primero no sea óptimo. Idealmente, antes de proponer un sistema de GAI, resultaría necesario cumplir otros criterios de ajuste más exigentes.

ABSTRACT

Analysis of the difficulty of a mathematics test using a componential model. The components involved in solving a test of mathematics are defined. The test include 66 items of operations with whole (positive and negative) numbers. The fit to Rasch model is studied and the difficulty parameters of components are estimated, using the Fischer LLTM model. 8 of 10 components were retained. A high linear relation was found between both difficulty parameters (r=0.8783). The results of this prior work are understood as a first promising step in order to design an automatic item generation system.

Key words: Rasch model, componential models, test of mathematics, automatic item generation.

REFERENCIAS

Alonso Tapia, J. y Olea, J. (1997). Modelos de evaluación de los conocimientos matemáticos. En J. Alonso (Dir.). *Evaluación del conocimiento y su adquisición*. Madrid: CIDE.

Assessment Systems Corporation (1989): RASCAL: *Rasch item calibration program*. St. Paul, MN: Autor.

Barnes, L. L. B. y Wise, S. L. (1991). The utility of a modified one-parameter IRT model with small samples. *Applied Measurement in Education, 4(2)*, 143-157.

Bejar, I. I. (1986). *Adaptive assessment of spatial abilities*. Research report. Princeton: Educational Testing Service.

Bejar, I. I. (1988). A sentence-based automated approach to the assessment of writing: A feasibility study. *Machine-Mediated Learning, 2*, 321-332.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. En N. Frederiksen, R. J. Mislevy e I. Bejar (eds.), *Test Theory for a New Generation of Tests*. NJ: LEA.

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium*. Research report. Princeton: Educational Testing Service.

Cuesta, M. (1996). Unidimensionalidad. En J. Muñiz (Coor.). *Psicometría*. Madrid: Universitas.

Dickson, L., Brown, M. y Gibson, O. (1991). *El aprendizaje de las matemáticas*. Madrid: Labor-MEC. (original de 1984).

Embretson, S. E. (1993). Psychometric models for learning and cognitive processes. En N. Frederiksen, R.J. Mislevy e I. Bejar (eds.), *Test theory for a new generations of tests.* Hillsdale, NJ: LEA.

Embretson, S. E. (1995). Developments toward a cognitive design system for psychological tests. En D. J. Lubinski y R. V. Dawis (eds.), *Assessing Individual Differences in Human Behavior. New Concepts, Methods and Findings.* Palo Alto: Davies Black.

English, L. y Halford, G. (1995). *Mathematics education: models and processes*. Hillsdale, NJ: LEA.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Fischer, G.H. (1997). Unidimensional Linear Logistic Rasch Model. En W.J.Van Der Linden y R.K. Hambleton (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Fischer, G. H. y Ponocny-Seliger, E. (1997). *LPCM-WIN Program*. Groningen: IEE. ProGAMMA

Fuson, K. C. (1992). Research on whole number addition and substraction. En D. Grows (ed.), *Handbook of Research on Mathematics Teaching and Learning*. New York: Mc Millan.

Grows, D.A.(1992). *Handbook of research on mathematics teaching and learning*. New York: Mc Millan.

Hambleton, R. K.; Swaminathan, H. y Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. CA: Sage Publications.

Hornke, L. F. y Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10(4)*, 369-380.

Irvine, S. H., Dann, P. L. y Anderson, J. D. (1990). Towards a theory of algorith-determined cognitive test construction. *British Journal of Psychology, 81*, 173-195.

Leder, G. (1992). *Assessment and learning of mathematics*. Howthorn, Vic.: Australian Council for Educational Research.

Marshall, S.P.(1993). Assessment schema knowledge. En N. Frederiksen, R.J. Mislevy e I. Bejar (Eds.). *Test theory for a new generation of tests*. Hillsdale, NJ: LEA.

Maza, C. (1989). *Sumar y restar: El proceso de enseñanza-aprendizaje de la suma y la resta*. Madrid: Visor.

Medina-Díaz, M.(1993). Analysis of cognitive structure using the Linear Logistic Test Model and Quadratic Assignment. *Applied Psychological Measurement*, 17, 2, 117-130.

Meisner, R., Luecht, R. y Reckase, M. (1993). *The comparability of the statistical characteristics of test items generated by computer algorithms*. Research report. Iowa: ACT research report series

Prieto, G. y Delgado, A. (1999). Medición cognitiva de las aptitudes. En J. Olea, V. Ponsoda y G. Prieto (Eds.), *Tests Informatizados: Fundamentos y aplicaciones*. Madrid: Pirámide.

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

Revuelta, J. y Ponsoda, V. (1998). Un test adaptativo informatizado de análisis lógico basado en la generación automática de ítems. *Psicothema, 10 (3)*, 709-716.

Revuelta, J. y Ponsoda, V. (1999). Generación automática de ítems. En J. Olea, V. Ponsoda y G. Prieto (Eds.), *Tests Adaptativos Informatizados: Fundamentos y aplicaciones*. Madrid: Pirámide.

Schwarz, B.B., Kohn, A.S. y Resnick, L.B. (1992). Bootstrapping mental constructions: A learning system about negative numbers. *ITS 92 Proceedings of the Second International Conference on Intelligence Tutoring Systems*. Montreal.

Sloboda, J.A. y Rogers, D. (1987). *Cognitive processess in mathematics*. Oxford: Oxford University Press.

Solano-Flores, G. (1993). Item structural properties as predictors of item difficulty and item association. *Educational and Psychological Measurement, 53*, 19-31.

Van Der Linden, W.J.y Hambleton, R.K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

(Revisión aceptada: 4/5/99)