

EVALUACIÓN DE LA COMPETENCIA MATEMÁTICA AL FINAL DE LA EDUCACIÓN PRIMARIA COMBINANDO TEORÍA CLÁSICA DE RESPUESTA AL TEST Y TEORÍA DE RESPUESTA AL ÍTEM

Rubén Fernández Alonso
E. O. E. P.- General del Sector Nalón
Manuel Rovés, 3, 1º. 33900. Ciaño (Langreo)
Tel.: 985680499; Fax: 985697032
e-mail: rubenfa@educastur.princast.es

RESUMEN

Este escrito presenta los datos más destacados que, hasta el momento, arroja el proyecto de estudio *Mathematikoi* cuyo objetivo es evaluar el nivel de competencia curricular en el área de matemáticas en la educación primaria. En el estudio se combinan los estadísticos derivados de la Teoría Clásica de Respuesta al Test (TCT) con los más modernos análisis basados en la Teoría de Respuesta al Ítem (TRI). Precisamente por emplear esta doble metodología la comunicación se estructura en dos apartados. En el primero se relata el proceso de creación y selección de una batería de preguntas que conforman un test baremado con los rudimentos de la TCT. La segunda parte está dedicada a la TRI. Dado que esta es la primera vez que en Asturias se emplea esta metodología para evaluar la competencia curricular en un nivel de planificación como es el sector educativo se hace una breve introducción a la misma. A continuación se presentan los resultados más destacados derivados del ajuste de la batería de preguntas seleccionadas con TCT a los modelos matemáticos basados en TRI. El trabajo finaliza apuntando algunas implicaciones prácticas del uso de la TRI en la evaluación de la competencia curricular.

1. INTRODUCCIÓN

En el curso 2000-01 el E.O.E.P.-General del Nalón (Asturias), dentro de su plan de formación interna, inició un estudio destinado a elaborar un sistema de evaluación y reeducación en el área de matemáticas para los tres ciclos de educación primaria. Lo que se presenta a continuación es una pequeña parte de ese trabajo. Concretamente el estudio evaluativo *Mathematikoi-III* realizado en el tercer ciclo de primaria.

La finalidad del proyecto *Mathematikoi* es elaborar un banco de ítems estandarizado que permita realizar evaluaciones objetivas de la competencia matemática. En la depuración de los ítems del banco se han combinado los estadísticos propios de la Teoría Clásica de Respuesta al Test (TCT) con los modelos matemáticos de la Teoría de Respuesta al Ítem (TRI). La presente comunicación tiene por objeto describir el proceso de selección y análisis de ítems fruto del cual se ha destilado un banco original de 38 ítems, que permite estimar el nivel de competencia curricular en el área de matemáticas, tanto al final de la educación primaria, como al inicio de la secundaria obligatoria.

2. ANÁLISIS DESDE TEORÍA CLÁSICA DE RESPUESTA AL TEST (TCT)

El trabajo se inició con el análisis del currículo de tercer ciclo del área de matemáticas: decretos de mínimos y curricular de Educación Primaria. También se revisaron los documentos elaborados por el MEC (Cajas Rojas, Propuestas de Secuenciación, etc.), libros de texto de las principales editoriales y el material de refuerzo y recuperación disponible en la sede del EOEP. Estas tareas sirvieron para diseñar la tabla de especificaciones de la prueba y construir un primer banco de ítems.

El banco primitivo de ítems consta de 500 preguntas y problemas que fueron valoradas con un doble objetivo. Por un lado, traducirlas a un formato de fácil contestación y corrección: preguntas cerradas con cuatro alternativas de respuesta. Por otro, asignar cada cuestión a una entrada de la matriz de especificaciones. Esta segunda tarea era bastante comprometida ya que con ella se busca asegurar la validez de contenido de la escala. La vocación curricular de la prueba exige que los ítems representen los objetivos y contenidos más relevantes del área. Finalmente, del medio centenar de preguntas se seleccionaron 60 que formaron parte del primer estudio piloto.

2.1. Estudio piloto

La tabla N.º1 presenta la **matriz de especificaciones** diseñada para el pilotaje de ítems.

TABLA N.º1. Distribución de los ítems en la matriz de especificaciones de la prueba piloto de 6º Educación Primaria

		CAPACIDADES MATEMÁTICAS				TOTAL	
		Conceptos	Algoritmos	Estrategias	Problemas		
BLOQUES DE CONTENIDO	Numeración	Enteros	2	2	****	****	13 (21,6%)
		Decimales	3	2	****	****	
		Fracciones	3	1	****	****	
	Operaciones	Enteros	0	3	2	2	20 (33,3%)
		Decimales	1	3	1	1	
		Fracciones	1	4	2	0	
	Geometría	Ángulos y Rectas	1	1	1	1	11 (18,3%)
		Geometría plana	1	0	1	1	
		Cuerpos-Volumen	1	1	1	1	
	Medidas	Longitud	2	1	1	0	12 (20%)
		Capacidad y Masa	1	2	1	0	
		Superficie	2	1	0	1	
	Organización de la información		1	1	1	1	4 (6,6%)
	TOTAL		19 (31,6%)	22 (36,6%)	11 (18,3%)	8 (13,3%)	60 (100%)

Como se observa la matriz tiene un total 46 entradas y se organizó en **dos ejes principales**: bloques de contenido y capacidades matemáticas. Estos dos grandes ejes de ordenación incluyeron diferentes categorías. De esta forma, las categorías contempladas dentro del eje **bloques de contenido** son –con ligeras modificaciones- las consideradas en los decretos curriculares: *Numeración*, *Operaciones*, *Geometría*, *Medidas* y *Organización de la información*. Salvo esta última, todas las demás también se dividieron en otras subcategorías. Así, los bloques *Numeración* y *Operaciones* distinguían entre números naturales, decimales y fraccionarios. El bloque *Geometría* se subdividió en: ángulos y rectas, geometría plana y cuerpos-volumen. Por último, en el apartado *Medidas* las subcategorías fueron: longitud, superficie, capacidad y masa.

Por otra parte, la tabla de especificaciones contemplaba cuatro **capacidades matemáticas**, entendidas éstas como “las operaciones cognitivas implicadas en la realización de los distintos tipos de tareas matemáticas” (INCE, 1998, 32). Estas capacidades son: 1/ *Conceptos*, es decir, conocimientos básicos (nombrar y reconocer objetos, conocer el sistema numérico decimal y otros sistemas convencionales de medida, etc.) 2/ *Algoritmos*: suponen, además de un conocimiento básico, el uso de una rutina sencilla (v.g. operación elemental de cálculo) o de un instrumento (v.g. de una regla, un transportador, etc.) 3/ *Estrategias intermedias*: implican la utilización de

varios algoritmos (p.ej., preguntas que combinan dos o más operaciones básicas) y la aplicación de alguna fórmula matemática (p.ej., el cálculo de áreas, superficies o perímetros). 4/ *Resolución de problemas* donde, además de todo lo anterior, se valora la capacidad de análisis, deducción, toma de decisiones, comparación, verificación, etc.

Las 60 preguntas fueron distribuidas al azar a **tres formas-cuestionario**. Por tanto, cada versión del cuestionario constaba de 20 preguntas. La aplicación fue colectiva a grupos-aula. Además, en el momento de la aplicación, las tres versiones piloto se distribuyeron azarosamente dentro de los grupos. Esta segunda aleatorización permite comparar resultados descartando que las posibles diferencias se deban a la dificultad intrínseca de las tres formas-piloto. La muestra estuvo compuesta por 6 grupos-aula de 1º de ESO de centros de EP e IES. Para evitar sesgos debidos a la edad –la prueba final está destinada al alumnado que finaliza la educación primaria- la aplicación se llevó a cabo al inicio del curso, cuando apenas habían transcurrido 25-30 días lectivos. Se excluyeron del análisis de resultados los cuestionarios contestados por los alumnos con necesidades educativas especiales. La tabla N.º2 presenta la muestra distribuida por el número de alumnos que contestaron a cada una de las formas-cuestionario.

TABLA N.º2. Distribución de la muestra por cuestionario

Forma A	32
Forma B	32
Forma C	25
Total	89

En el registro de los datos se consideró: 1, acierto y 0, error. Se reservó el 9 para los porcentajes de ns/nc. En el análisis estadístico se empleó el programa SPSS V.8.0. **Los 60 ítems fueron seleccionados siguiendo dos criterios básicos**. Uno de ellos fue el estadístico. Dentro de éste se consideraron, a su vez, dos indicadores: el índice de discriminación y el de dificultad de cada ítem. En principio se eliminaron todos los ítems con un índice de discriminación (correlación ítem-test) inferior a 0,2. El índice de dificultad (porcentaje de aciertos) se tuvo en cuenta en todos los ítems ya que en la prueba final se pretendía incluir preguntas fáciles, difíciles y de dificultad media. El segundo criterio para la depuración de ítems fue la aportación de los mismos a la validez de contenido de la prueba. Este segundo criterio fue más restrictivo que la propia selección estadística ya que dejó fuera ítems con buenos índices numéricos pero que aportaban poco a la selección muestral de contenidos.

Finalmente, de los 60 ítems pilotados se seleccionaron 34. A estos 34 se unieron, de cara a la aplicación final, 7 preguntas nuevas hasta completar una prueba de 41 ítems. Aunque la inclusión de preguntas no probadas en el pilotaje inicial implica cierto riesgo no supone una gran violación metodológica puesto que todos los ítems sufrirán una segunda depuración estadística. Y, en todo caso, era conveniente asumir dicho riesgo ya que los nuevos ítems apuntalaban la selección muestral del contenido, a la vez que permitían un tiempo de aplicación no superior a la hora.

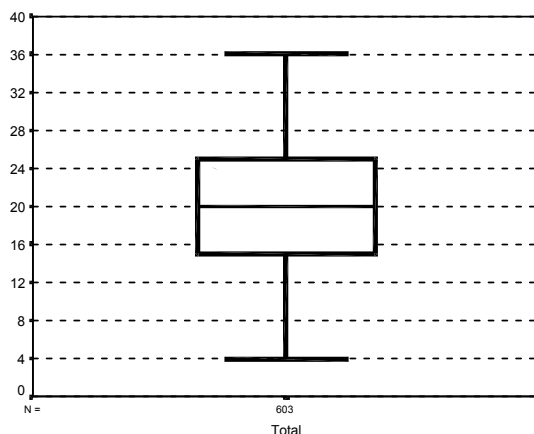
2.2. Estudio final

La aplicación final, compuesta por la batería de 41 preguntas, se llevó a cabo en última semana de mayo y primera de junio de 2001. El universo de población se definió como el alumnado escolarizado en 6º de educación primaria en el Sector Educativo del Nalón

en el curso 2000-01. Según los datos del EOEP-Nalón, el número de estudiantes en ese nivel era de 693. La muestra final que tomó parte en el estudio fue de 603 estudiantes (de ellos, el 51,7% son mujeres y el 66,8% estudia en un centro público). Por tanto, se evaluó al 87% del alumnado previsto en el parámetro poblacional. Esto permitió trabajar con un **error muestral muy pequeño ($\pm 1,5\%$)** que autoriza –con un nivel de confianza del 95%- a generalizar los resultados, cuanto menos, a todos los escolares del valle del río Nalón. En el estudio participaron finalmente 26 colegios. De éstos, 18 (69,2%) son de titularidad pública y 8 privados-concertados. En total se evaluaron 37 unidades: 24 (64,8%) públicas y 13 concertadas.

En el registro de datos y análisis de ítems se siguió el mismo procedimiento y programa estadístico que en el estudio piloto. También fueron idénticos los criterios de selección de ítems. En este caso se eliminaron tres ítems. El primero por tener un índice de discriminación prácticamente nulo. Un segundo ítem se cribó por presentar un índice de discriminación pequeño (aunque superior a 0,2) y aportar poco a la selección de contenido. La causa de la última eliminación fue un error de formulación en los distractores detectado en el análisis cuantitativo de respuestas. Dicho error tenía como consecuencia la carga desmesurada de las respuestas hacia una alternativa falsa. Al despreciar estas tres preguntas los cálculos de tipificación y fundamentación estadística del protocolo final *Mathematikoi-III* fueron realizados sobre los 38 ítems restantes. La baremación (cálculo de percentiles y puntuaciones típicas *Z*) fue posible gracias al buen comportamiento general de la escala. Aunque se ha calculado todo el conjunto de estadísticos previstos desde la Teoría Clásica sirva de ejemplo **dos gráficos para demostrar cómo la tipificación ha sido altamente satisfactoria.**

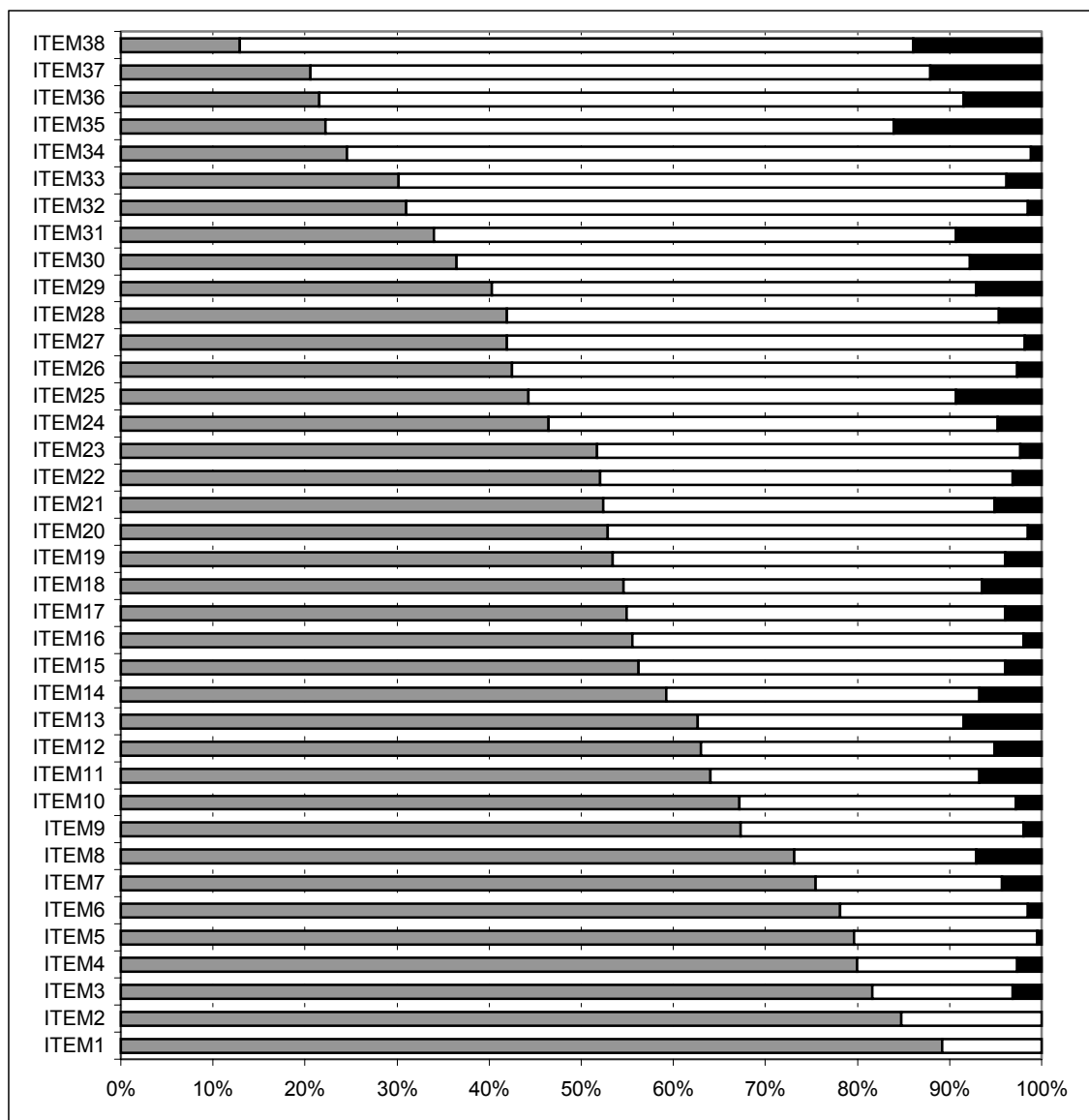
GRÁFICO N.º1. Representación de la distribución de los resultados mediante gráfico de caja



El diagrama de caja del gráfico N.º1 ofrece una vista cenital de la distribución de resultados. La altura de la caja representa la amplitud intercuartil, es decir, el lugar que ocupa el 50% de la muestra. El borde superior de la caja es el percentil 75 y el inferior corresponde al percentil 25. La línea central es la mediana (percentil 50). Los límites al final de la línea vertical son las puntuaciones máximas y mínimas. Como se puede ver, el percentil 75 corresponde a 25 puntos directos en la prueba y el percentil 25 a 15 puntos directos. Por último, la puntuación máxima alcanzada fue de 36 puntos y la mínima 4. Si se observa el gráfico con detenimiento es posible comprobar que el diagrama de caja tiene una simetría perfecta. Es decir, del percentil 25 al 50 hay la misma distancia (5 puntos directos) que desde el percentil 50 al 75. Además, la distancia desde la puntuación mínima a la mediana (16 puntos directos) también es idéntica a la distancia entre la mediana y la puntuación máxima. Estas proporciones del

gráfico de caja son un excelente indicador de las posibilidades de tipificación y baremación de la escala.

GRÁFICO N.º2. Ítems de la escala *Mathematikoi-III* ordenados por su índice de dificultad



Un segundo dato que habla de las posibilidades de tipificación del instrumento es la distribución del índice de dificultad de los ítems. En el gráfico N.º2 se presenta el porcentaje de aciertos de los 38 ítems de la escala *Mathematikoi-III*. El área gris de la barra representa el porcentaje de acierto; el área blanca el porcentaje de error; y el negro el ns/nc. El gráfico permite apreciar la progresividad del porcentaje de acierto ítem a ítem y demostrar cómo la escala está en condiciones de graduar rendimientos.

Para cerrar el apartado dedicado a la presentación de datos desde Teoría Clásica se adjuntan los siguientes **estadísticos de fundamentación** de la prueba:

- Índice de discriminación: 0,39 (media de las correlaciones ítem-test)
- Índice de dificultad: 0,52 (proporción media de aciertos)
- Índice de fiabilidad: 0,86 (calculado con el alpha de Cronbach)

- Índice de validez: 0,70 (correlación entre nota otorgada con anterioridad por el tutor en matemáticas y el resultado en la prueba)

Para comprobar su bonanza, los datos se comparan con los conseguidos por el Instituto Nacional de Calidad y Evaluación (INCE) en sus trabajos sobre evaluación de la competencia curricular en primaria y secundaria obligatoria.

TABLA N.º3. Comparación de los índices de fiabilidad de la escala *Mathematikoi-III* con los resultados encontrados por el INCE en la evaluación del nivel curricular de la educación primaria

	Promedio en Pruebas Piloto	Prueba Final
Mathematikoi-III	0,79	0,86
Matemáticas 6º EP	0,83	0,85
Matemáticas 2º EP	0,84	0,86
Lengua 6º EP	0,80	0,89
Lengua 2º EP	0,85	0,88
Conocimiento del medio 6º EP	0,81	0,87

FUENTE: Elaboración propia a partir de INCE (1997): *Evaluación de la educación primaria*, Madrid, MEC, 203.

Como se puede ver en la primera tabla la fiabilidad de la escala *Mathematikoi-III* es similar a las logradas en las evaluaciones del INCE en las áreas de Matemáticas, Lengua y Conocimiento del medio, tanto en términos absolutos, como en ganancia entre el pilotaje y el estudio final.

TABLA N.º4. Comparación de los índices de fiabilidad, dificultad y discriminación de la escala *Mathematikoi-III* con los resultados encontrados por el INCE en la evaluación del nivel curricular de la secundaria obligatoria

	Fiabilidad	Dificultad	Discriminación
Mathematikoi-III	0,86	0,52	0,39
Matemáticas 14	0,87	0,43	0,39
Matemáticas 16	0,89	0,48	0,42
Literatura y reglas lingüísticas 14	0,85	0,52	0,38
Literatura y reglas lingüísticas 16	0,85	0,51	0,38
Comprensión lectora 14	0,84	0,59	0,49
Comprensión lectora 16	0,81	0,67	0,35
Ciencias de la Naturaleza 14	0,72	0,36	0,20
Ciencias de la Naturaleza 16	0,73	0,37	0,21
Geografía e Historia 14	0,86	0,46	0,29
Geografía e Historia 16	0,83	0,46	0,26

FUENTE: Elaboración propia a partir de INCE (1998): *Los resultados escolares*. En *Diagnóstico del sistema educativo. La escuela secundaria obligatoria*.1997, Madrid, MEC, 131.

En la segunda tabla se comparan los índices de fundamentación estadística del *Mathematikoi-III* con las pruebas de evaluación de la secundaria obligatoria. El propio INCE califica de “altamente satisfactorios” sus resultados. Como se puede ver la fiabilidad de *Mathematikoi-III* no desmerece a ninguna de las baterías de evaluación del INCE. Por su parte, el índice de dificultad raya el promedio teórico de 0,5 lo que no ocurre con las pruebas de Ciencias Naturales, que resultaron difíciles, o con la prueba de Comprensión Lectora-16 años que resultó demasiado fácil. Por último, el índice de discriminación de *Mathematikoi-III* resiste bien la comparación con las pruebas del INCE –exceptuando Comprensión Lectora-14 años, que fue más discriminante que el

resto- e incluso se sitúa bastante por encima de algunas escalas como las de Ciencias de la Naturaleza y Geografía e Historia.

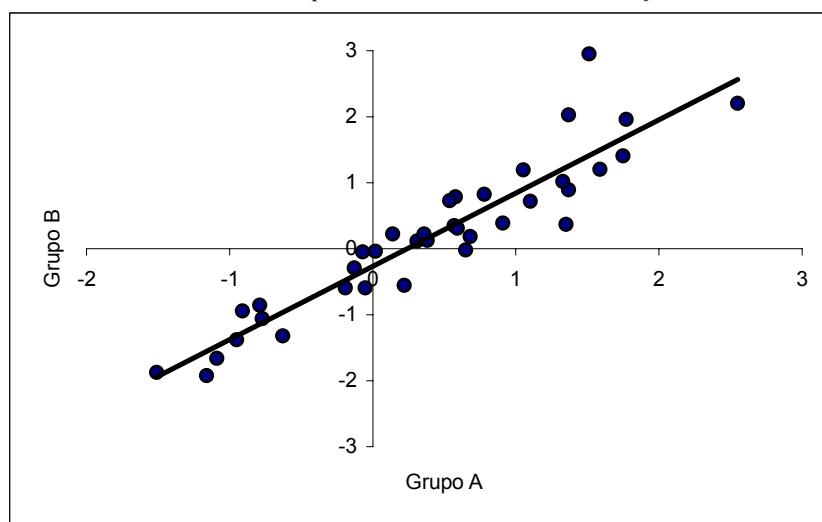
3. ANÁLISIS DESDE TEORÍA DE RESPUESTA AL ÍTEM (TRI)

3.1. Breve presentación de la Teoría de Respuesta al Ítem

En una segunda fase del estudio los 38 ítems de la prueba *Mathematikoi-III* fueron calibrados con modelos basados en TRI. Dado que este estudio aplica por primera vez en Asturias dicha teoría a la evaluación curricular desde un sector educativo no está de más hacer, a modo de presentación, una breve reseña de este nuevo enfoque de investigación.

El enfoque clásico tiene, sin duda, una gran tradición y es el empleado por la mayoría de los evaluadores y constructores de test. Sin embargo, la TCT adolece de una serie de limitaciones tanto teóricas como metodológicas. Intentando superar estos límites surge la TRI. No es intención repasar las ventajas de la TRI, por otro lado descritas desde las cátedras de psicometría (Martínez-Arias, 1995; Muñiz, 1997). Aquí sólo mencionará lo que se considera **la principal aportación de la TRI frente al enfoque clásico: ofrecer puntuaciones invariantes con respecto al instrumento empleado y a la muestra con la que se estandariza el test**. Cuando se barema un test desde TCT el resultado depende de la muestra. Es decir, el test será fácil si los sujetos son muy competentes y difícil si los sujetos son menos aventajados. De igual forma, si las preguntas son fáciles los resultados serán mejores que si las preguntas son difíciles. La TRI supera este inconveniente al utilizar una escala de medida indeterminada. Es decir, una escala cuyo rango de puntuación no va desde 0 a la puntuación máxima del test (en el caso de *Mathematikoi-III* de 0 a 38), sino una escala donde el rango va de $-\infty$ a $+\infty$. Con ello se logra romper la dependencia del instrumento y de la muestra de referencia en la interpretación de resultados. Un ejemplo gráfico quizá aclare mejor este concepto básico.

GRÁFICO N.º3. Invarianza del parámetro b en dos muestras de diferente rendimiento

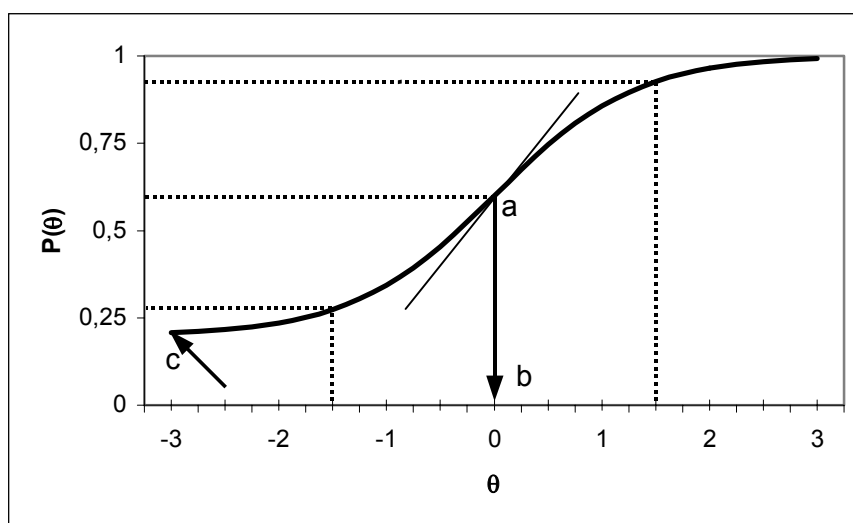


El gráfico bivariante presenta datos reales del estudio *Mathematikoi-III*. Se seleccionan dos muestras (Grupos A y B) cuyo rendimiento en la prueba difiere estadísticamente. Es decir, hay diferencias significativas entre los resultados de los dos grupos. A

continuación, se calcula el índice de dificultad de los 38 ítems en cada uno de los grupos. Como se verá más adelante, cuando se trabaja en el marco de la TRI el índice de dificultad se denomina parámetro b y su valor numérico no es el mismo que en el enfoque clásico. En este punto debe recordarse que en TCT si se encuentran dos muestras con diferencias significativas de rendimiento es necesario realizar dos baremaciones: una para cada grupo. En cambio, como se aprecia en el gráfico N.º3, al usar los procedimientos TRI el parámetro de dificultad se mantiene constante independientemente del grupo que se elija para su estimación. Esto se comprueba calculando la correlación entre los índices de dificultad de los dos grupos. En este caso la correlación es de 0,93. Se trata de una correlación prácticamente perfecta que indica que la prueba arroja similares resultados en los grupos aunque éstos difieran en su nivel de competencia. Es decir, la estimación de la dificultad del ítems es independiente del nivel de conocimientos de la muestras.

Una vez presentada la que se considera principal ventaja en la interpretación de resultados de la TRI se pasa a exponer de forma muy somera, por las limitaciones de espacio, **la idea central** de este nuevo enfoque de la medida. La TRI se sustenta sobre la suposición de que **existe una función matemática que conecta la capacidad o el nivel de una persona en una o más variables –que por convención se denomina nivel θ – con la probabilidad de acertar, $P(\theta)$, un ítem concreto diseñado para evaluar dicha capacidad. Esta función matemática se denomina Curva Característica del Ítem.** Cada ítem tiene su propia curva característica, es como su código de identificación. Dicha curva, como se viene diciendo, predice la probabilidad de acertar un ítem, $P(\theta)$ a un determinado nivel θ según sea ese ítem, es decir, según se definan una serie de parámetros. Aunque son muchos los modelos matemáticos que pueden dar cuenta de esta suposición, en el proyecto *Mathematikoi* los análisis se han realizado usando los modelos logístico-unidimensionales. Más concretamente los resultados que se presentan en esta comunicación están derivados del modelo logístico de tres parámetros. Se llama modelo logístico porque utiliza la función logaritmo para describir la conexión entre el nivel θ y la probabilidad de acierto, $P(\theta)$. Es unidimensional porque intenta medir una única variable, en este caso la competencia matemática. Y se denomina “de tres parámetros” porque son tres los indicadores o rasgos que sirven para definir las propiedades métricas de cada ítem.

GRÁFICO N.º4. Curva Característica del Ítem



En el gráfico N.º4 se presenta un ejemplo ficticio de Curva Característica. En el eje de coordenadas se encuentra el valor de la variable, θ , cuyo rango va desde $-\infty$ a $+\infty$, aunque en este ejemplo el rango de puntuación sólo abarque de -3 a +3. El eje de ordenadas, $P(\theta)$, representa la probabilidad de acertar un ítem y, como toda probabilidad, su valor oscila entre 0 y 1. En el presente ejemplo un sujeto con un nivel de habilidad $\theta = 0$ tiene una probabilidad de 0,60 de acertar el ítem. La probabilidad de acierto de un sujeto con $\theta = -1,5$ es 0,27 y la probabilidad de acierto de un sujeto con $\theta = 1,5$ es 0,92. Es decir, a mayor nivel en θ mayor $P(\theta)$. En el modelo de tres parámetros las Curvas Características se definen por tres rasgos o parámetros del ítem, que también se recogen en el gráfico:

- **Parámetro a.** También llamado índice de discriminación porque, como en Teoría Clásica, hace referencia a la capacidad discriminante del ítem. Sin embargo, su valor no es el mismo que en TCT ya que cuando la variable sigue la distribución normal –media 0 y desviación típica 1– el valor de a es: $a \equiv r_b \sqrt{1-r_b^2}$, donde r_b es la correlación ítem-test, el índice de discriminación clásico. La cuantía numérica del parámetro a es proporcional a la inclinación de la recta tangente a la curva característica en el punto de máxima pendiente de ésta.
- **Parámetro b** o índice de dificultad del ítem. Es el valor de θ para el punto de máxima pendiente de la curva característica. Su valor numérico no es igual que en el enfoque clásico. Cuando θ es normal $N(0,1)$ b es: $b \equiv -Z_p / r_b$, donde Z_p es la puntuación típica del índice de dificultad clásico y r_b la correlación ítem-test. Una característica de este parámetro es que está expresado en la misma escala de medida que θ y, por tanto, sus valores también van de $-\infty$ a $+\infty$.
- **Parámetro c.** Es la probabilidad de acertar un ítem al azar, es decir, cuando no se conoce la respuesta. Expresa el valor de la $P(\theta)$ cuando ésta tiende a $-\infty$.

3.2. Principales resultados e implicaciones prácticas.

En este último epígrafe se presentan algunos de los resultados más importantes de la aplicación del modelo logístico de tres parámetros a los datos del estudio *Mathematikoi*. En todo trabajo desde la óptica TRI las **dos tareas fundamentales** son la comprobación del modelo y la calibración de ítems.

En la fase de **comprobación del modelo** es necesario verificar los supuestos asumidos desde la TRI y la invarianza de parámetros. El principal supuesto se denomina unidimensionalidad, que significa que el banco de ítems a calibrar mide una única variable. El modo habitual de comprobar este supuesto es mediante análisis factorial. En este caso se encontró un primer factor dominante que explicaba algo menos del 20% de la varianza de resultados. Aunque, desde la ortodoxia académica, este índice puede parecer pequeño trabajos como los de Reackase (1979); Cuesta (1996); Muñiz y Cuesta (1993) permiten concluir que es posible aplicar modelos unidimensionales a los datos del estudio *Mathematikoi*. Como ya se advirtió al presentar la TRI la invarianza de parámetros es la principal ventaja frente a la TCT. En aquel momento se ofreció el dato de la invarianza del parámetro b, que era estable para dos muestras con diferencias significativas en el rendimiento. De igual forma se ha comprobado la invarianza en el parámetro θ , es decir, en la habilidad estimada de los estudiantes. En este caso el modo

de verificar la invarianza es elegir dos grupos de ítems con índices de dificultad diferentes –un grupo de ítems fáciles y otro difíciles- y comprobar que la estimación del nivel θ de los alumnos no varía en ninguno de los dos grupos de ítems. Se han encontrado correlaciones entre el nivel θ estimado con ítems fáciles y difíciles superiores a 0,8. Se trata de un índice que, sin ser tan bueno como el de la invarianza del parámetro de dificultad, permite afirmar la existencia de una alta asociación (y, por tanto, invarianza) entre la estimación del rendimiento del alumnado cuando contesta a los ítems fáciles o a los difíciles.

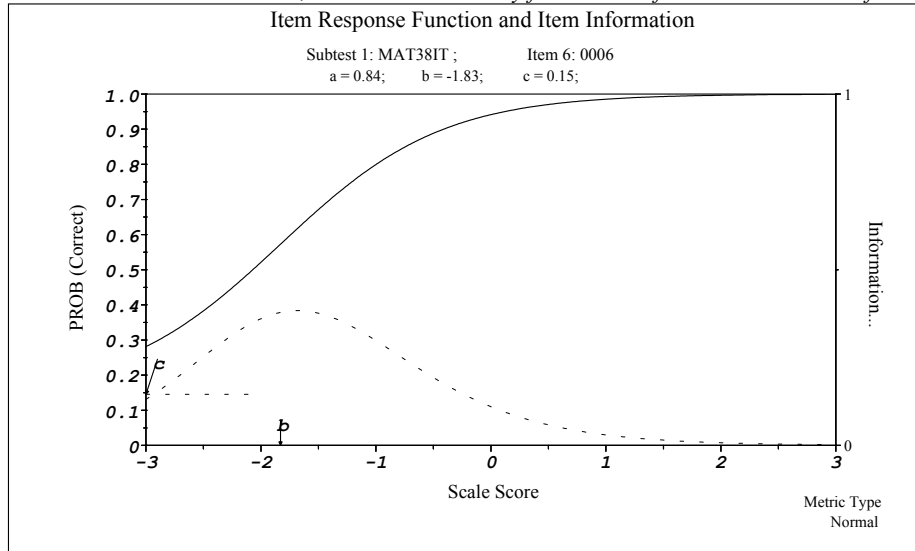
Comprobado el modelo el segundo paso es **calibrar los ítems**, es decir, calcular sus parámetros, comprobar su ajuste al modelo elegido y estimar el nivel θ del alumnado. En este análisis se empleó el programa BILOG 3.0 que permite realizar las calibraciones de ítems a cualquiera de los 3 modelos logísticos para variables dicotómicas. Todos los ítems, salvo uno, ajustaron al modelo de tres parámetros con una probabilidad $\alpha = 0,01$. Por ello es posible concluir que, en su conjunto, la batería *Mathematikoi-III* ajusta al modelo elegido y, por tanto, que se dispone de un banco de ítems capaz de evaluar la competencia matemática al final de la primaria o al inicio de la secundaria obligatoria. Además, por las propiedades matemáticas de la TRI los 38 ítems configuran un banco inicial, susceptible de incrementarse en el futuro con nuevos ítems.

Llegado a este punto es posible preguntarse cuál es la **ganancia en términos prácticos de la TRI frente a la TCT**. Es decir, qué implicaciones educativas se derivan del empleo de la TRI. De forma breve la respuesta sería que un banco de ítems calibrado desde TRI **permite llevar a cabo evaluaciones rápidas y adaptadas a las características de los alumnos**, cuestión imposible en el enfoque clásico. En TCT el núcleo del análisis es el test, la prueba objetiva y estandarizada. Cuando se evalúa la competencia de una persona con una escala baremada desde TCT es necesario aplicar el protocolo completo para estimar su nivel de rendimiento. En cambio, desde la óptica TRI no es necesario aplicar un banco de ítems completo: con implementar unos pocos ítems se tiene la evidencia necesaria para estimar la habilidad de los evaluados. Esto es así porque en TRI el protagonismo pasa del test al ítem. Ya no interesa el test en su conjunto, sino el ítem individual. La tarea clave no es estandarizar una prueba, sino estimar los parámetros de los ítems y aprehender sus propiedades métricas, definiéndolos mediante su curva característica. Así que al disponer de un banco de ítems calibrado es posible calcular el nivel θ de los evaluados a partir de un conjunto pequeño y representativo de ítems de dicho banco. Pero, además de rápidas, la TRI permite realizar evaluaciones adaptadas a las características de los sujetos. Esto es posible gracias a que la TRI ofrece un tratamiento más novedoso y, a la vez, plausible del error de medida. En TCT el intervalo de confianza del error de medida es constante a lo largo de la escala. Sin embargo, en TRI se establece un intervalo confidencial para cada uno de los valores θ . Es decir, un ítem determinado no tiene la misma precisión para todos los niveles de la escala θ . Esto supone que dentro de un banco de ítems los hay más adecuados para medir en ciertos niveles θ . Precisamente para aquellos donde el error estándar es más pequeño. Esto es lo que se conoce como función de información del ítem. Todos los ítems tienen una función de información determinada para cada nivel θ . Y hay un punto de la escala donde la cantidad de información del ítem es máxima. Pues bien, para ese punto de θ es para el que el ítem es más adecuado.

Veamos con datos extraídos del estudio *Mathematikoi* como ocurre esto. A modo de ejemplo se presentan tres ítems con diferentes parámetros y funciones de información.

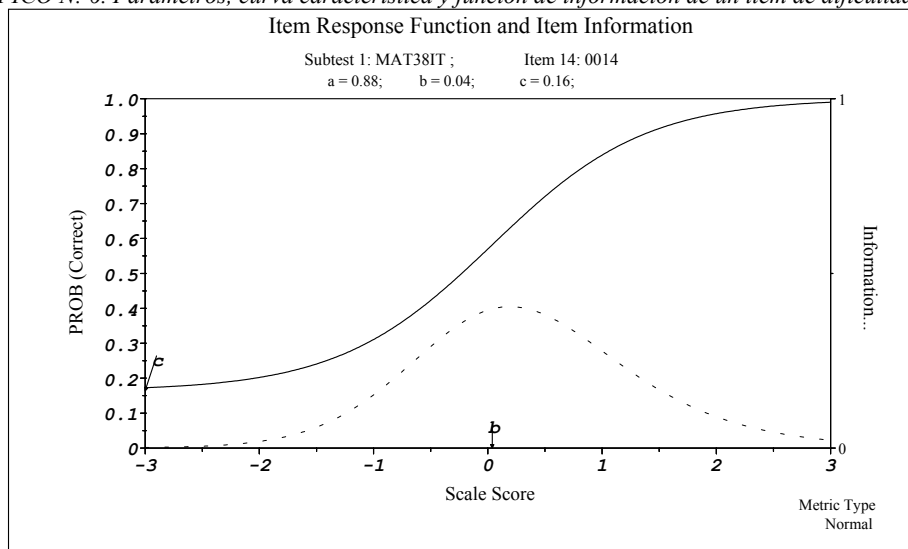
Las salidas gráficas son las que ofrece el programa BILOG V. 3.0. La línea sólida es la curva característica y la punteada la función de información. En el título, debajo del nombre del test y del número de identificación del ítem, aparecen los valores de los tres parámetros.

GRÁFICO N.º5. Parámetros, curva característica y función de información de un ítem fácil



El primer gráfico representa un ítem discriminante ($a = 0,84$), muy fácil ($b = -1,83$) y con una probabilidad de acierto por azar de $0,15$. Analizando la curva característica se aprecia que la probabilidad de acierto para alumnos poco competentes es relativamente alta (aproximadamente $0,3$ para $\theta = -3$) y muy alta (más de $0,9$) para estudiantes con conocimientos medios $\theta = 0$. Por su parte, la línea de la función de información es más alta para el alumnado con puntuaciones θ entre -3 y -1 . Sin embargo, no es un buen ítem para discriminar entre rendimientos medios y altos, ya que una vez θ toma valores positivos la altura de la curva que representa la función de información del ítem se vuelve despreciable.

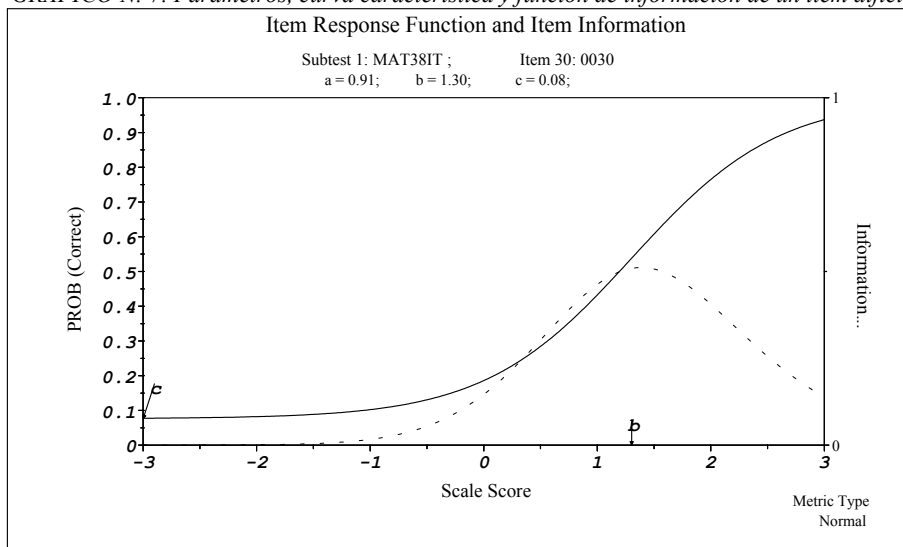
GRÁFICO N.º6. Parámetros, curva característica y función de información de un ítem de dificultad media



El segundo gráfico representa un ítem con similar poder discriminante y probabilidad de acierto por azar a la pregunta anterior (es este caso, $a = 0,88$ y $c = 0,16$). Sin embargo,

ahora se está ante un ítem de dificultad media ($b = 0,04$). Según la curva característica del ítem la probabilidad de acierto esperada para los alumnos menos competentes (nivel θ menor que -1) es muy baja ($0,2$). Por el contrario, aquellos cuyo nivel θ es mayor de $1,5$ tienen probabilidades de acierto superiores a $0,9$. La curva que representa la función de información del ítem confirma estos comentarios. El ítem ofrece más información, es decir, es más preciso y discriminante en los valores θ comprendidos entre $-0,5$ y $+1$.

GRÁFICO N.º7. Parámetros, curva característica y función de información de un ítem difícil



Por último, se presenta un ítem de dificultad media-alta ($b = 1,3$). Se trata de un ítem con poder discriminante ($a = 0,9$) y baja probabilidad de ser acertado al azar ($c = 0,08$). Su curva característica indica que no discrimina entre rendimientos bajo y medio. La probabilidad de acierto para $\theta = -3$ es menor de $0,1$ y para $\theta = 0$ no supera el $0,2$. En cambio, permite discriminar los rendimientos medio-alto y alto. Así, la probabilidad de acierto para $\theta = b$ ($1,3$) es ligeramente superior a $0,5$ y sólo el alumnado con $\theta = 2$ o superior tienen una probabilidad mayor de $0,8$ de responder acertadamente el ítem. Una vez más la función de información confirma estas impresiones. El ítem es más preciso, es decir, tiene un error de medida más pequeño para los valores θ positivos y más concretamente para aquellos comprendidos entre $0,5$ y 3 .

Las consecuencias prácticas de esto son muy importantes para los evaluadores y constructores de pruebas objetivas. Con un banco de ítems calibrado es posible seleccionar aquellos que mejor se ajusten al nivel de los sujetos. Centrándose en la evaluación de la competencia curricular se elegirán ítems fáciles o muy fáciles para estimar el nivel de los alumnos con mayores dificultades y necesidades educativas. Por el contrario, con alumnos talentosos se seleccionarían aquellos ítems que ofrecen máxima información en los niveles θ más altos. Al contar con un banco calibrado, el E.O.E.P.-Nalón ha iniciado la construcción de pruebas de evaluación rápidas adaptadas a diferentes niveles de rendimiento en matemáticas. Incluso, el mercado ya ofrece programas como el MICROCAT que, además de calibrar ítems y estimar niveles θ , permite realizar este tipo de evaluaciones de forma automatizada. Por ello, puede ser de gran ayuda al profesorado no familiarizado con los procedimientos de cálculo de la TRI. Sin embargo, hablar de las posibilidades de la informática en la evaluación curricular supera el objetivo y el espacio de esta comunicación.

INCE (1996): **Lo que aprenden los alumnos de 12 años. Evaluación de la educación primaria. Datos básicos 1995.** INCE/MEC, Madrid.

INCE (1997): **Evaluación de la educación primaria.** INCE/MEC, Madrid.

INCE (1998): *Los resultados escolares.* En **Diagnóstico del sistema educativo. La escuela secundaria obligatoria. 1997,** INCE/MEC, Madrid.

INCE (2000): **Evaluación de la educación primaria. Datos básicos 1999.** INCE/MEC, Madrid.

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