

# ANÁLISIS DE ÍTEMS PARA LA EVALUACIÓN DE LA COMPETENCIA CURRICULAR EN EL ÁREA DE MATEMÁTICAS

RUBÉN FERNÁNDEZ ALONSO y  
RAQUEL FERNÁNDEZ RODRÍGUEZ\*

El presente artículo describe los resultados de un estudio piloto, destinado a realizar una selección de ítems cara a construir protocolos definitivos de evaluación de la competencia curricular en el área de matemáticas, al final del primer ciclo de la Educación Primaria. El trabajo recrea todas las fases del pilotaje, desde el diseño inicial de la tabla de especificaciones, hasta el cribado definitivo de 140 ítems que fueron reducidos a 57. La conclusión final es que el proceso de selección ha resultado satisfactorio, lo que permite esperar que en un futuro se dispondrá de instrumental fiable, discriminante, ajustado a la normalidad y válido desde el punto de vista del contenido, para evaluar los objetivos terminales en el área de matemáticas al finalizar el segundo curso de Educación Primaria.

The article describes the processes carried out to build a preliminary performance test in mathematics to be applied to students of the first levels of primary school. The authors describe the steps followed to build the test: from the initial design of the table of specifications to the selected 57 items of the 140 initial ones. As a result a satisfactory test has been produced, which will allow to increase the number of available reliable, discriminative, valid and standardized techniques written in Spanish to assess students' performance in mathematics at the end of the second year of primary school.

## 1. Justificación

En el curso 2000-01 el Equipo de Orientación Educativa y Psicopedagógica del Nalón (EOEP-Nalón) se constituyó en grupo de trabajo, iniciando una serie de estudios cuya finalidad es evaluar la competencia curricular en el área de Matemáticas en Educación Primaria. Es decir, desde el nivel de final de primer ciclo de Educación Primaria, hasta el inicio de la Secundaria Obligatoria. El conjunto de estos estudios fue bautizado con el nombre genérico de *Proyecto Matematikoi*, y recibe financiación por parte de la Viceconsejería de Educación del Principado de Asturias.

Parece haber argumentos sólidos para defender la pertinencia de los estudios que se están realizando. La reforma estructural que supuso la LOGSE partió en dos etapas la escolaridad obligatoria. Ello motivó un gran número de cambios organizativos y curriculares. La orientación escolar no fue ajena a estas novedades. Primero porque, beneficiándose del peso reconocido en la nueva ordenación, se implantaron los departamentos en los Institutos de Enseñanza Secundaria (IES). Además, fue necesario replantear

---

\* RUBÉN FERNÁNDEZ ALONSO y RAQUEL FERNÁNDEZ RODRÍGUEZ son Psicopedagogos del Equipo de Orientación Educativa y Psicopedagógica del Nalón, Asturias.

sus funciones en los niveles donde ya estaba institucionalizada. No es lo mismo orientar en una etapa con carácter final, como era la antigua EGB, que en una fundamentalmente propedéutica como es la Educación Primaria. De esta forma las tareas académico-vocacionales, que tenían un peso fuerte en los Servicios de Orientación Educativa y Vocacional (SOEV), perdieron gran parte del mismo.

Por otra parte, incluso antes de la reforma del 90, la investigación educativa básica había llegado a cierto consenso sobre la evidencia de que *el mejor predictor del rendimiento académico es el propio rendimiento*. Es decir, que el conocimiento previo de los alumnos es la base de todo proceso de enseñanza. La información curricular se convierte en una variable imprescindible para ponderar desempeños académicos. Llega al nivel de consideración de las áreas clásicas de evaluación psicométrica-cognición y personalidad. Ahora bien, en los albores de la reforma parecen faltar instrumentos que midan la competencia curricular.

Las carencias de la medida del nivel académico no son una problemática uniforme. Tradicionalmente la literatura se ha preocupado más por el área de lengua. Gracias a ello se dispone de instrumental fiable y variado para hacer estimaciones en amplios bloques de contenido del lenguaje: velocidad y comprensión lectura o precisión y expresión escrita. Sin embargo, en otras áreas las deficiencias de material para evaluar el rendimiento son mayores. Por ello, se vio la necesidad de integrar en el modelo predictivo una medida fiable del rendimiento en Matemáticas. Esta afirmación tampoco impide reconocer la existencia de material de evaluación en otras áreas. Así, por ejemplo, ya antes de la reforma se había publicado la obra de Arnal (1988). En la actualidad existe un conjunto de pruebas objetivas para estimar conocimientos escolares. Algunas centradas en el área de Matemáticas como la obra de Barbero, I. y Navas, M.J. (1995). También otras diseñadas como baterías de evaluación de diferentes áreas de conocimiento. Ejemplos de este último tipo de trabajos son los de Alonso Tapia (1997). El trabajo propone la evaluación de cuatro áreas en Secundaria Obligatoria: Ciencias Sociales (vol. I), Ciencias Naturales (vol. II) y Matemáticas y comprensión lectora (vol. III). Por último, para la etapa de Primaria, Pérez Avellaneda y otros (1999), presenta cuatro escalas: madurez, lenguaje, matemáticas y conocimiento del medio.

Así mismo, existe toda una pléyade de test comerciales clásicos que evalúan el factor de inteligencia general a través de la estimación de subfactores entre los que generalmente se incluye el razonamiento matemático o el cálculo numérico. Ejemplos de este tipo de baterías son:

- La serie TEA de Thurstone y Thurstone (1979), que mide tres factores básicos (V, R, N), el último de los cuales es específicamente numérico.
- La escala de Bennett y otros (1985), que al igual que el anterior mide factor verbal, razonamiento abstracto y numérico. Pero en este caso las tareas propuestas no se centran sólo en el cálculo o la solución de problemas. También recoge preguntas sobre sistemas

- de medida convencionales y geometría, lo que confiere a la escala un marcado carácter curricular.
- También son destacables dos obras de Yuste (1985 y 1995). En ambas se busca estimar un Factor G mediante la evaluación de diferentes subfactores entre los que se encuentran la aptitud numérica y para el cálculo.
  - Para la etapa de Educación Infantil y primer ciclo de Educación Primaria, es bien conocida la prueba de Boehm (1982) que entre sus 50 conceptos básicos pregunta por algunos relativos a los números y la cantidad. También es muy popular la prueba de McCarthy (1983), factorializada tripartitamente en las escalas verbal, manipulativo-preceptiva y numérica. Esta última escala, que intenta evaluar la "aptitud" numérica, está compuesta por tres pruebas: cálculo, memoria numérica y recuento y distribución. Por último, la batería de Yuste y Sanz (1991) evalúa hasta 102 conceptos básicos que se ordenan en tres categorías fundamentales: espaciales, temporales y cuantitativo/numéricos. Dentro de esta última categoría distingue a su vez conceptos relativos a: contar, añadir, quitar, repartir, comparar cantidades y magnitudes y discriminar numerales y símbolos aritméticos.

Explicar por qué, en comparación con el área de Lengua, existen menos trabajos sobre evaluación del aprendizaje en Matemáticas traspasa los límites de la presente exploración de datos. Por ahora nos contentaremos con apuntar que en la década de los 80 la investigación evaluativa en el área experimenta un punto de inflexión. En el año 1986, el National Council of Teachers of Mathematics (NCTM) comienza los trabajos destinados a elaborar un conjunto de indicadores del currículo de Matemáticas de carácter comprensivo, también conocidos como los estándares P-12. Existe una traducción de dichos estándares realizada por la Sociedad Andaluza de Educación Matemática (1991).

Por esas mismas fechas, y con un enfoque de investigación más empírico, la Asociación Internacional para la Evaluación del Rendimiento Educativo (IEA) realizaba sus primeros estudios transnacionales de evaluación de matemáticas y ciencias. España participó por primera vez en el estudio: **Evaluación Internacional de la Mejora del Rendimiento** (IEAP) de 1988 donde concurren 5 países y 4 provincias canadienses (Canadá no posee un sistema federal de educación). Se puede encontrar una versión en castellano de este trabajo, donde no aparece informe técnico, sólo presentación y discusión de resultados, en Lapointe, Mead y Phillips (1989).

Por lo que respecta a nuestro país, en la década de los 90 el Instituto Nacional de Calidad y Evaluación (INCE) comenzó sus estudios destinados a medir el nivel de competencia curricular en la Educación Primaria y Secundaria Obligatoria (INCE, 1996, 1997, 1998 y 2000). Con ellos se demostraron las posibilidades de la medición masiva del nivel curricular. El

proyecto *Mathematikoi* se inspira en esta segunda tradición de trabajos buscando diseñar pruebas de evaluación fiables, robustas y con varias aplicaciones. Algunas de las cuales se enumeran a continuación.

El uso más directo del instrumento se centra en la evaluación individual. La identificación de posibles dificultades y/o altas capacidades en el área. La escala propone un abanico de tareas que permite evaluar no sólo conocimientos, sino también procedimientos y estrategias de solución de tareas matemáticas.

Por otra parte, la batería está siendo tipificada y se podrá aplicar colectivamente en una sesión de 50 minutos. La finalidad de la escala cualitativa es traducir las puntuaciones obtenidas en la prueba a contenidos y capacidades matemáticas. Por ello, puede ayudar en la definición de objetivos y contenidos de adaptaciones curriculares o simplemente a determinar los próximos avances que potencialmente hará el alumnado.

Pero no todas las utilidades tienen que ver con recolectar y traspasar información. Este tipo de instrumental ha tenido múltiples usos. De esta forma la puntuación de la escala –entendida como la operativización de la variable “conocimientos previos en Matemáticas”– puede formar parte de un modelo de predicción académica más amplio. De hecho, en el marco de las aplicaciones colectivas del EOEP-Nalón, e incluso con muestras pequeñas (unos 20 alumnos) hemos encontrado correlaciones significativas entre los resultados en las escalas *Mathematikoi* y otras medidas curriculares tales como el rendimiento en lectura (velocidad y comprensión), ortografía y el cálculo mental. Todo ello parece indicar una buena validez concurrente de nuestras escalas con otros test que estiman diferentes áreas del rendimiento académico. A su vez, el conjunto de variables curriculares se puede contrastar con otros constructos como habilidades intelectuales, motivación y expectativas, estilos de aprendizaje, etc., hasta agotar un hipotético modelo de diferencias individuales. En este sentido, los análisis factoriales realizados parecen apuntar al hecho de que las variables curriculares se agrupan en un único factor, que, a su vez, es independiente del factor general de inteligencia.

Los fundamentos estadísticos de las escalas *Mathematikoi* también permiten utilizarlas en la evaluación de programas de enseñanza matemática, definición de estándares de centro, presentación de datos a los alumnos y familias, toma de decisiones, etc.

La situación actual del proyecto *Mathematikoi* es la siguiente: se han realizado tres estudios piloto y otros tantos finales. Los números I, II y III de los estudios *Mathematikoi* coinciden con las edades de baremación de cada prueba, 8, 10 y 12 respectivamente. Es decir, final de 1º, 2º y 3º ciclo de primaria.

En total han participado más de dos mil alumnos y se han probado cientos de ítems. La tabla 1 recoge las fechas y las distribuciones de número de alumnos e ítems pilotados y logrados por aplicación.

Tabla N.º 1: Fecha y distribución del número de alumnos e ítems pilotados y logrados por aplicación

	Fecha de las aplicaciones	Ítems pilotados	Ítems logrados	Muestra
Mathematikoi-III (estudio piloto)	Octubre 2000	60	34	89
Mathematikoi-III (estudio final)	Junio 2001	41	38	603
Mathematikoi-II (estudio piloto)	Octubre 2001	100	58	212
Mathematikoi-II (estudio final)	Junio 2002	58	57	490
Mathematikoi-I (estudio piloto)	Octubre 2002	140	57	128
Mathematikoi-I (estudio final)	Junio 2003*	60	-	500

## 2. Objeto y Objetivo

No es intención del presente escrito describir el proyecto Mathematikoi al completo. El objeto del análisis es más modesto, ya que sólo se centrará en una pequeña parte del proyecto. En concreto, se intentará recrear el diseño y los resultados del estudio piloto correspondiente a 2º de Educación Primaria. Es decir, a partir de ahora el trabajo se circunscribirá exclusivamente en el pilotaje Mathematikoi-I, cuya aplicación final se encuentra en la actualidad en proceso de elaboración.

El objetivo del estudio piloto Mathematikoi-I es probar el funcionamiento de 140 preguntas y reducir las mismas hasta seleccionar 60 ítems aproximadamente. Es decir, la predicción inicial era conseguir alrededor del 40% de los reactivos pilotados. Fijar de antemano un número de ítems a considerar (60) se hizo pensando en el producto final buscado. Se pretendía, de cara a la aplicación final, diseñar dos pruebas de, aproximadamente, 30 ítems cada. La longitud de reactivos por prueba se hizo sobre la estimación de que ése era el número razonable de preguntas que los estudiantes de 2º de Primaria podrían contestar en menos de una hora.

Al construir dos cuestionarios, se dispondrá de dos pruebas, validadas desde la Teoría Clásica de los Tests, para evaluar el rendimiento en Matemáticas al final del primer ciclo de Primaria. Pero además, en el

\* Los datos de la aplicación final Mathematikoi-I son previsiones para junio de 2003.

análisis de resultados de la aplicación final también se empleará la metodología de la Teoría de Respuesta al Ítem. Así que las dos pruebas tendrán una serie de preguntas comunes con las que se creará, después de la aplicación final, un único banco de ítems por medio de un diseño de anclaje. Por tanto, la previsión final es disponer de dos test baremados clásicamente y de un único banco de ítems ajustado a Teoría de Respuesta al ítem.

### **3. Contexto Poblacional y Muestra Seleccionada**

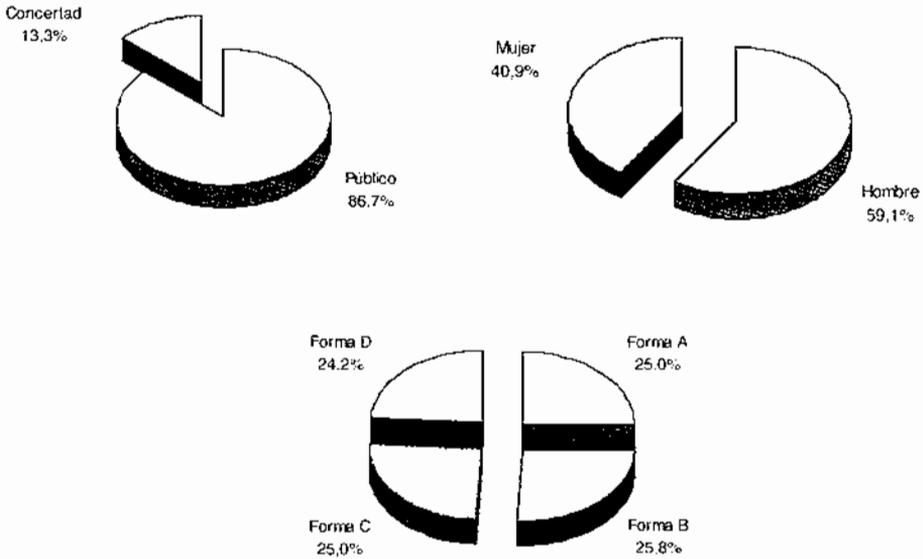
El sector educativo del Nalón está ubicado en la zona central de Asturias, abarcando parte del valle por el que discurren los tramos medio y alto del río Nalón. Administrativamente se encuentra constituido por los concejos de Langreo, San Martín del Rey Aurelio, Laviana, Sobrescobio y Caso, agrupados en la Mancomunidad de Municipios del Valle del Nalón. No es un sector educativo muy grande puesto que la extensión geográfica total es de 645,48 kilómetros cuadrados.

Sus características socioculturales conforman un tipo de población determinada. En principio, y con carácter general, es necesario apuntar que el nivel sociocultural de las familias corresponde con estudios primarios y/o secundarios (Bachillerato o Formación Profesional). Por su parte, en muchos casos, los ingresos familiares, es decir, el nivel socioeconómico de las familias se derivan directa o indirectamente del trabajo en las minas y puede considerarse con medio o medio-alto. También existen un porcentaje considerable de minorías étnicas. De hecho, en algún centro 1 de cada 4 estudiantes pertenece a minoría étnica. En los últimos años también se han comenzado a escolarizar hijos de inmigrantes, aunque la proporción de los mismos es aún reducida.

Sobre esta población se seleccionó la muestra del estudio. Al final participaron 9 centros públicos y concertados. En total se evaluaron 128 alumnos de 3º de Educación Primaria del sector educativo del Nalón. El muestreo de los centros se hizo de la siguiente manera: cada uno de los psicopedagogos del EOEP-Nalón eligió de modo aleatorio un centro de entre los que tiene asignado. De igual forma, en el caso de que el centro dispusiera de más de una unidad en el nivel elegido, se seleccionó una de ellas para participar en el estudio. En algún caso fue necesario que algún miembro del EOEP seleccionara una segunda unidad para lograr una muestra de, al menos, 30 estudiantes por forma pilotada con el fin de garantizar el supuesto de normalidad en la distribución de resultados.

Los siguientes gráficos de pie recogen la distribución del porcentaje de estudiantes por tres variables de clasificación: sexo, titularidad del centro y forma-cuestionario a la que contestaron.

Gráfico N° 1: *Distribución del porcentaje de estudiantes por sexo, titularidad del centro y forma contestada.*



Como se puede ver en el primer gráfico, el porcentaje de hombres es sensiblemente superior al de las mujeres. Éste es un problema de selección muestral puesto que la distribución de este porcentaje en el nivel de 3° de Educación Primaria del sector educativo es coincidente con los parámetros de la población estatal. Es decir, hay una distribución paritaria en función del sexo.

Por su parte, también ha sido mayoritario el número de estudiantes escolarizados en centros públicos, frente a los concertados. Esto ocurre también en el sector educativo, donde los alumnos matriculados en centros públicos es superior a los concertados. Sin embargo, la proporción para el total del sector es de 65% de alumnos en centros públicos y 35% en concertados.

Por último, la distribución de los alumnos por el cuestionario al que se contestó ha sido prácticamente idéntica en las 4 formas. Con ello, se pone de manifiesto que el proceso de aleatorización de las formas dentro de los grupos aula se ha llevado a cabo satisfactoriamente.

#### 4. Diseño y Procedimiento

En este punto se describirá el estudio piloto *Mathematikoi I*, correspondiente a 2° de Educación Primaria. Como ya se apuntó

anteriormente la finalidad del pilotaje es realizar una selección de ítems para, posteriormente, construir y validar estadísticamente una prueba que permita estimar el nivel curricular de los alumnos en el área de Matemáticas al final del primer ciclo de la Educación Primaria.

El trabajo se inició con el análisis del currículo de primer 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 contenidos de la prueba y construir un primer banco de ítems.

La tabla de contenidos (tabla 2) 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*.

Sin duda, en el primer ciclo de Primaria el uso de números y operaciones es el contenido más importante y, por ello, al que se dedica más tiempo. Es, además, “la capacidad más característica del ámbito de las matemáticas” (MEC, 1992). En cuanto a los **números**, el decreto de enseñanzas mínimas propone el conocimiento del sistema decimal de numeración hasta la centena y el valor posicional de unidades, decenas y centenas.

Por lo que se refiere a las **operaciones**, las orientaciones ministeriales aconsejan centrarse básicamente en sumas y restas, dejando abierta la posibilidad de iniciarse en la multiplicación en contextos adecuados. Una vez más, la observación de los niveles de competencia del alumnado permiten afirmar que al finalizar el primer ciclo hay un porcentaje significativo de alumnos que conocen el algoritmo de la multiplicación e, incluso, un selecto grupo de privilegiados que se han iniciado en los rendimientos de la división. El rendimiento de estos alumnos está muy por encima de las expectativas de la ordenación académica ya que con respecto a la resta, se afirma que, dada su complejidad, será necesario profundizar en las restas con llevadas en el segundo ciclo de Primaria.

Se aconseja, así mismo, el destacar el papel del cálculo mental en el desarrollo de la capacidad de operar (suma y resta) con números, así como el apoyo que la estimación de resultados o el cálculo aproximado suponen para la consolidación de una operatoria eficaz por parte del alumnado.

En cuanto al bloque de **sistemas de medidas**, el legislador recomienda el desarrollo de la capacidad de medir, muy unida a la necesidad de comparar, lo que supone promover en una primera fase la búsqueda de estrategias e instrumentos no convencionales que satisfagan esa necesidad para, posteriormente, ir introduciendo unidades e

instrumentos progresivamente más convencionales y eficaces. Así, será conveniente introducir unidades de tiempo (año, día, hora), de longitud (metro, centímetro), de capacidad (litro) o de masa (kilogramo), para propiciar el inicio en el uso de instrumentos de medida convencionales sencillos (regla, reloj, balanza doméstica de lectura directa). Así mismo, se hace hincapié en la importancia de desarrollar la capacidad de estimar resultados ligados a la medida, haciendo énfasis en la utilidad de las estimación, así como en la necesidad de elegir tanto instrumentos como unidades adecuadas a aquello que se quiere medir.

En cuanto al bloque de contenido de **geometría**, se considera una característica de este ciclo la capacidad de reconocer en el entorno y en los objetos que en él se encuentran distintos cuerpos y formas geométricas (esfera, cubo, cilindro, círculo, rectángulo, triángulo) "que ayuden a entenderlo mejor, a describirlo y a comunicarse con los demás acerca de él" (MEC, 1992).

A lo largo del primer ciclo de Primaria, también se considera conveniente empezar a desarrollar estrategias de **organización de la información**, a través de distintas actividades, para que sean capaces de registrar un suceso (recuento, agrupación), representarlo (tablas, gráficos sencillos) y comunicarlo a los demás (lectura, comprensión, expresión) o utilizar su información para otras actividades. En este sentido, las orientaciones ministeriales prevén que los contenidos de este bloque ayuden a desarrollar las siguientes capacidades:

- Realización de representaciones gráficas sobre situaciones y hechos muy concretos y que sean capaces de interpretarlas.
- Comprender y expresar a su manera una situación matemática que se les proponga, es decir, que entiendan el enunciado y las consignas dadas.
- Favorecer el desarrollo de estrategias personales de resolución y organización de datos necesarios de problemas, siendo conveniente que el alumnado exprese oralmente las decisiones que han ido tomando en cada momento.
- Desarrollar la capacidad de persistir en la exploración de un problema.

La configuración del segundo eje de clasificación, **capacidades matemáticas**, fue más problemática y discutida. Esto se debió tanto a razones epistémicas, como de definición operativa de las propias categorías. Sin embargo, la inclusión de este segundo eje era una tarea insoslayable a la vista de los últimos desarrollos en este campo de la investigación educativa. Hace más de una década que los estudios transnacionales de la IEA consideran las capacidades matemáticas dentro de sus estimaciones de rendimiento (Lapointe, Mead y Phillips, 1989). Más recientemente han comenzado a ser observadas en las evaluaciones oficiales españolas (INCE, 1996, 1997, 1998 y 2000). Se entiende por capacidades matemáticas las destrezas necesarias para contestar a una pregunta concreta. El INCE (1998,

23) las define como las “operaciones cognitivas implicadas en la realización de los distintos tipos de tareas matemáticas” que incluyen cuestiones que van “desde los conceptos básicos, hasta la aplicación de los conceptos y algoritmos en la resolución de problemas complejos de la vida cotidiana”.

En la tabla de contenidos de la prueba piloto se han contemplado tres capacidades matemáticas: Conocimientos básicos, Algoritmos y Resolución de problemas.

- Se entiende por *Conocimientos básicos* la adquisición de una serie de conceptos fundamentales. Se incluyen representar, nombrar o reconocer objetos y propiedades (p.ej., ángulos, figuras geométricas, etc.), dominar el sistema numérico decimal y otros sistemas convencionales de medida (seriaciones, redondeos, unidades básicas de medida, etc.).
- Los *Algoritmos* implican a los conceptos pero van más allá. Suponen, además, el uso de una rutina sencilla (una operación elemental de cálculo) y pueden incluir también las actividades que demuestran el uso de instrumentos (una regla para medir segmentos o un transportador para medir ángulos).
- Por último, en la Resolución de problemas se incluyen la utilización de estrategias intermedias que implican el uso 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), y también las tareas que implican las capacidades de análisis, deducción, toma de decisiones, comparación, verificación, etc.

Tabla N.º 2: Distribución de número y porcentaje de ítems pilotado en la tabla de contenidos.

		Capacidades			
		Conceptos	Algoritmos	Problemas	Total
Bloques de contenido	Números-operaciones	15 (10,7%)	38 (27,1 %)	27 (19,3 %)	80 (57,1 %)
	Geometría	21 (15,0%)	5 (3,6 %)	4 (2,8 %)	30 (21,4 %)
	Medidas	9 (6,4%)	5 (3,6 %)	2 (1,4 %)	16 (11,4 %)
	Organización información	3 (2,1%)	5 (3,6%)	6 (4,3 %)	14 (10,0 %)
	Total	48 (34,3%)	53 (37,8 %)	39 (27,8 %)	140 (100 %)

En la tabla 2 se presenta la matriz de especificaciones que sirvió de instrumento para la selección muestral del contenido. Como se puede ver, la tabla dispone de 12 entradas. Para cubrir dichas entradas se elaboraron las 140 preguntas pilotadas. Eran preguntas cerradas, es decir, de opción múltiple, con tres o cuatro alternativas de respuesta y sólo una verdadera. En cada entrada de la tabla aparece el número de ítems que se le asignó y, entre paréntesis, el porcentaje que, con respecto al total, le corresponde a

cada casilla de la tabla. Así, por ejemplo, la entrada correspondiente a “conceptos de números-operaciones” engloba 15 preguntas, que equivalen al 10.7% del total de ítems pilotados. En los marginales, el porcentaje no siempre corresponde con la suma de las filas o las columnas. Esto se debe a los redondeos realizados para ofrecer el porcentaje con un solo decimal.

El peso de cada una de las entradas, así como de cada una de las categorías que conforman los dos ejes de clasificación (contenidos y capacidades), se estableció mediante un criterio de expertos. La ponderación otorgada intenta ser el reflejo que, en el desarrollo curricular, tiene cada uno de los contenidos y capacidades evaluadas.

Se estimó en más de 200 minutos el tiempo necesario para contestar las 140 preguntas. Ello obligaba a dividir en varias sesiones la aplicación de la prueba piloto con los inconvenientes que esto acarrea: mayor riesgo de fatiga en los estudiantes, menor operatividad del instrumento, más esfuerzo hasta completar una muestra con suficientes sujetos, etc. Con el fin de minimizar estas objeciones, se llevó a cabo un doble proceso de aleatorización.

Basándose en criterio de expertos, las 140 preguntas fueron ordenadas por índice de dificultad y distribuidas en cuatro formas-cuestionario. Por tanto, cada versión constaba de 35 preguntas. La aplicación fue colectiva a grupos-aula. Cada alumno sólo respondió a una de las cuatro formas-piloto. Los estudiantes emplearon una media de 40 minutos en completar el cuestionario y, salvo rarísimas excepciones, nadie tardó más de 50 minutos. Así se logró reducir el tiempo de aplicación y neutralizar los posibles efectos de la fatiga en el rendimiento. Además, en el momento de la aplicación, las cuatro versiones-piloto se repartieron aleatoriamente dentro de los grupos-aula. Esto permitió comparar resultados descartando que las posibles diferencias se debiesen a la dificultad intrínseca de las formas-piloto.

La aplicación piloto se lleva a cabo al inicio del curso 2002-03 con el fin de que los alumnos de 3º curso no hayan avanzado mucho en sus conocimientos curriculares y tengan un nivel de competencia bastante parecido al que se supone a los estudiantes que finalizan 2º de Primaria. Se excluyeron del análisis de resultados los cuestionarios contestados por los alumnos con necesidades educativas especiales.

Aplicados los cuestionarios, éstos son corregidos y posteriormente vaciados a una base de datos utilizando el programa SPSS V.10. En el registro de datos se consideró “1 = acierto” y “0 = error”. Se reservó el 9 para calcular los porcentajes de “no sabe/no contesta”.

## 5. Selección de Ítems

En la elección de ítems se consideraron dos criterios fundamentales:

1. Estadístico: los ítems deben atender a dos condiciones básicas. La primera, ser discriminantes, es decir, ser capaces de diferenciar entre los buenos y malos rendimientos. En segundo lugar, se busca que, en su conjunto, los ítems que conformarán cada cuestionario se ajusten a la normalidad en cuanto a la distribución de sus índices de dificultad.
2. Curricular: los ítems seleccionados deben ser pertinentes desde el punto de vista del contenido. Es necesario que después de la selección las preguntas representen del modo más amplio y válido con contenidos y capacidades matemáticas propias del final de primer ciclo de Primaria.

El primer criterio empleado para la eliminación de ítems fue el estadístico. En concreto se calcularon las correlaciones ítem-test de todos los reactivos y se eliminaron aquellos cuyo índice de discriminación no llegó a 0,20. Este criterio es el que emplea el Instituto Nacional de Calidad y Evaluación en el cribado de sus ítems y también es empleado en el proyecto *Mathematikoi* (FOEP-Nalón, 2000, Fdez. Alonso y otros 2003). En este caso, sólo 111 de los 140 ítems cumplieron esta condición. Sin embargo, también se incluyeron otros 6 reactivos cuyas correlaciones ítem-test se encontraban muy cercanas al punto de corte —entre 0,18 y 0,19— y se consideraban pertinentes desde el punto de vista curricular. Eran, pues, ítems que, sin alcanzar estrictamente el primer criterio estadístico, cumplían perfectamente la segunda condición.

En definitiva, al sumar los 111 ítems más discriminantes con los 6 resacados por su aportación a la validez de contenido, en la primera criba fueron salvados 117 ítems y cribados 23. Del total de ítems eliminados, 8 pertenecían al cuestionario-forma A; 6 a la forma B; 5 a la C; y 4 a la D. Con estas formas reducidas se recalcularon los índices de fiabilidad, dificultad y discriminación de cada cuestionario, así como algunos los estadísticos básicos (media, desviación típica, error estándar de media y rango) y se compararon con los resultados obtenidos antes de la eliminación de ítems con el fin de comprobar las posibles ganancias derivadas de la selección. Los resultados se presentan en la dos siguientes tablas.

En la primera se comparan algunos estadísticos básicos hallados antes y después del primer cribado. Atendiendo a la media (la cual debe ser interpretada en función del número de ítems) antes de la selección el cuestionario A era el más fácil seguido del B y el D, que presentaban valores similares. El cuestionario C parece el más difícil, puesto que la media hallada no llegaba a la media teórica, es decir, 17,5 puntos. Después de la eliminación de los primeros ítems, la situación se mantiene de modo similar, aunque las medias caen un poco en los cuestionarios A, B y C. En cambio, en el cuestionario D, aunque también desciende la media (de 18,5 a 17 puntos), en realidad esta caída no es tal, sino que supone una recuperación puesto que la media de la prueba, al quedar reducida ésta a 31 ítems, pasa a ser de 15,5 puntos.

Todos los cuestionarios, salvo la forma C, ven reducida la variabilidad de los resultados, es decir, la desviación típica. Esto se debe, entre otras cosas, a trabajar con un menor número de ítems. Al bajar la variabilidad, se encuentran disminuciones del error estándar de media en los cuestionarios A y B, manteniéndose el mismo error en los cuestionarios C y D. Sin embargo, la ganancia en la precisión de la media en las formas es muy pequeña, debido tanto a que las muestras no son amplias y a que la disminución de la varianza de resultados tampoco fue significativa.

Por último, el comportamiento de los rangos puede considerarse adecuado salvo en la forma C. En principio la forma D incluso aumentó su rango, pese a la disminución de ítems. En las formas A y B se perdieron dos puntos directos en el rango, pero como después de la selección se trabaja con 27 y 29 ítems respectivamente el porcentaje de rango cubierto pasó de 74,2% antes de la selección a 88,8% para la forma A y 82,8% para la forma B. En el extremo contrario se encuentra la forma C, que sufrió una importante disminución del rango de puntuaciones, cuestión no descable en este tipo de trabajos.

Tabla N.º 3. Comparación de algunos descriptivos básicos de los 4 cuestionarios antes y después de la primera selección estadística.

	ANTES DE PRIMERA SELECCIÓN					DESPUÉS DE PRIMERA SELECCIÓN				
	Nº Items	$\bar{X}$	$S_x$	$S_{\bar{x}}$	Rango	Nº Items	$\bar{X}$	$S_x$	$S_{\bar{x}}$	Rango
Forma A	35	19,5	5,5	1,0	26	27	14,4	5,3	0,9	24
Forma B	35	18,8	7,4	1,3	26	29	15,2	6,9	1,2	24
Forma C	35	16,7	5,8	1,0	25	30	13,9	5,8	1,0	19
Forma D	35	18,5	7,5	1,3	27	31	17,0	7,4	1,3	28
Total	140	18,4	6,6	0,6		117	15,1	6,3	0,5	

La siguiente tabla compara los índices de fiabilidad, dificultad y discriminación antes y después de la primera selección. Como comentario general se puede afirmar que los resultados han sido los esperados, aunque es necesario realizar algunas puntualizaciones.

En principio, todas las formas-cuestionario, salvo la D, han logrado ganancias en los índices de fiabilidad. Como coeficiente de fiabilidad se ofrece el alfa de Cronbach. La mayor ganancia se ha dado en la forma B, seguida de los cuestionarios A y B. En realidad, excepción hecha del

cuestionario D, los índices de fiabilidad de los cuestionarios después de la reducción de ítems son muy elevados si se tiene en cuenta tanto la naturaleza de las pruebas como la edad de los estudiantes.

Por su parte, no se aprecian cambios significativos en los índices de dificultad que se ha calculado como la proporción media de aciertos en cada forma. Es cierto que, con la eliminación de ítems, todo los cuestionarios, salvo el D, han rebajado su índice de dificultad. Es decir, todos se han hecho un poco más difíciles. Sin embargo, todos los guarismos de dificultad se mueven entre 0,50 y  $\pm 0,05$ , es decir, entre 0,45 y 0,55, que pueden considerarse índices de dificultad totalmente aceptables.

Por último, se puede comprobar las ganancias en los índices de discriminación de los cuestionarios, entendido como la media de las correlaciones ítem-test. La mayor ganancia se produce en la forma A (que pasa de 0,34 a 0,44); a continuación, con ganancias similares se encuentran los cuestionarios B, C y D.

Tabla N.º 4. Comparación de los índices de fiabilidad, dificultad y discriminación de los 4 cuestionarios antes y después de la primera selección estadística.

	ANTES DE PRIMERA SELECCIÓN			DESPUÉS DE PRIMERA SELECCIÓN		
	Fiabilidad	Dificultad	Discrimina.	Fiabilidad	Dificultad	Discrimina.
Forma A	0,78	0,55	0,34	0,84	0,53	0,44
Forma B	0,80	0,53	0,33	0,88	0,52	0,40
Forma C	0,79	0,47	0,36	0,83	0,46	0,42
Forma D	0,71	0,52	0,43	0,71	0,54	0,48

En definitiva, con esta primera etapa de eliminación de ítems se han logrado ganancias en los índices de discriminación y fiabilidad de los cuestionarios, mientras que el índice de dificultad se mantiene estable dentro de los parámetros esperados. Todo ello permite concluir que el primer descarte de ítems ha arrojado unos resultados satisfactorios.

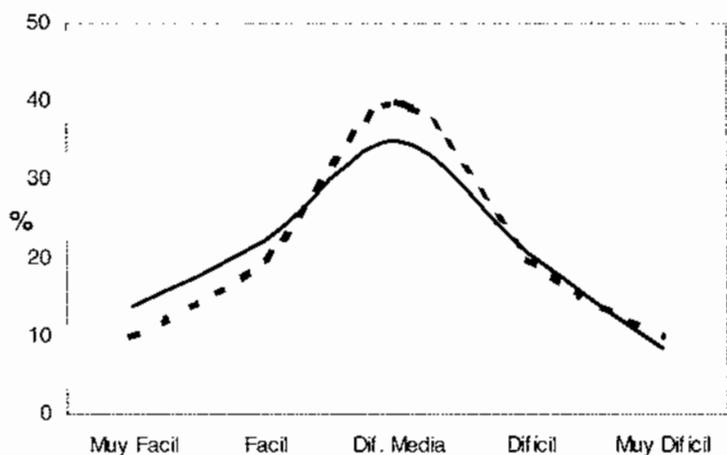
El segundo criterio de tipo estadístico para la selección de ítems es que éstos, en su conjunto, dispongan de índices de dificultad que se distribuyan de forma normalizada. Para lograr esto se crearon cinco categorías de ítems según su dificultad:

1. Ítems muy difíciles: menos del 19,99% de aciertos
2. Ítems difíciles: entre el 20 y el 39,99% de aciertos
3. Ítems de dificultad media: entre 40 y 60% de aciertos
4. Ítems fáciles: entre el 69,01 y el 80% de aciertos
5. Ítems muy fáciles: más del 80,01% de aciertos

Para asegurar la normalidad de resultados se establecieron a priori los siguientes pesos para cada una de las categorías: 10% para ítems muy fáciles y otro tanto para ítems muy difíciles, 20% de ítems fáciles y el mismo peso para ítems difíciles. Por último, 40% de ítems de dificultad media. Esta distribución teórica fue muy similar a la obtenida en el estudio empírico de los 117 ítems. En el siguiente gráfico se comparan la distribución teórica de los índices de dificultad de los ítems y la real de los ítems que han sobrevivido al primer análisis.

En el gráfico la línea punteada representa la distribución teórica de los índices de dificultad, mientras que la línea sólida indica el la distribución empírica. Como se puede ver, el dibujo es muy similar en ambas. La mayor discrepancia se encuentra en los ítems de dificultad media que suponen el 35% de los seleccionados, frente al 40% esperado. En todo caso, dado que esto es sólo una primera tentativa, hay que apuntar que los datos parecen muy satisfactorios.

Gráfico N.º 2: Comparación de la distribución teórica de los índices de dificultad con la distribución empírica de los 117 ítems de la primera selección.



Aunque los 117 ítems cumplen las dos condiciones estadísticas—ser discriminantes y presentar una distribución de sus índices de dificultad ajustados a normalidad— este número de reactivos es aún muy amplio. Lo que se busca es construir dos pruebas cuyo tiempo de aplicación no supere la hora en cada una de ellas. Se estima que cada cuestionario debería tener una longitud de 30 preguntas. Es decir, los 117 ítems deberían quedar reducidos a 60 aproximadamente.

En esta segunda fase es muy importante que los reactivos seleccionados representen los objetivos y contenidos del área en un peso

aproximado a la ponderación otorgada en la matriz de especificaciones original.

Sobre esta previsión de ítems, el porcentaje de los mismos según el peso de la tabla de contenidos debiera ser el siguiente:

Tabla N° 5: Distribución del porcentaje teórico de 60 ítems en la matriz de especificaciones.

	Conceptos	Algoritmos	Problemas	Total
<b>Números/Operaciones</b>	10 – 11,6	26,6	18,3 – 20	56,6
<b>Geometría</b>	15	3,3	1,6 · 3,3	20 – 21,6
<b>Medidas</b>	5 – 6,6	3,3	0 – 1,6	10 – 11,6
<b>Organización Información</b>	1,6	3,3	3,3 – 5	10
<b>Total</b>	33,3 – 35	36,3 – 38,3	26,6 – 28,3	100

Para lograr esta merma de ítems se siguió el siguiente procedimiento: se realizó un listado con las 117 preguntas, sus índices estadísticos, una breve descripción de los mismos y la entrada de la tabla a la que estaban asignados. El listado de ítems fue acompañado de la N.º5 de este escrito para tener presentes cuantos ítems eran necesarios por entrada de la matriz de especificaciones.

Con esta información se empleó un criterio de expertos que fueron analizando y seleccionando los ítems buscando la mayor relevancia curricular de los ítems a seleccionar. Así, a igualdad de índices de discriminación o dificultad se eligieron aquellos que mejor representaban el contenido y el tipo de tareas propias del currículo de matemáticas de 2º de primaria. Además, se consideró como criterio adicional que dentro de cada una de las entradas de la matriz de especificaciones también se respetara la distribución de normalidad de los índices de dificultad.

Con estas directrices se fueron destilando los reactivos hasta lograr una selección de 57 preguntas. No se llegaron a las 60 previstas en el primer momento porque posteriormente se incluirán 6-7 ítems más provenientes de la prueba de 2º ciclo. Esta inclusión de preguntas de la prueba de cuarto curso se hace para disponer de un diseño de anclaje que permita –mediante un procedimiento de equiparación- poner en la misma escala los resultados de primer y segundo ciclo. Dado que estos ítems provenientes de una prueba anterior no son objeto, en sentido estricto, de la presente selección, no serán considerados ahora. Por ello la nueva tabla sólo considera las 57 preguntas provenientes del estudio piloto de primer ciclo.

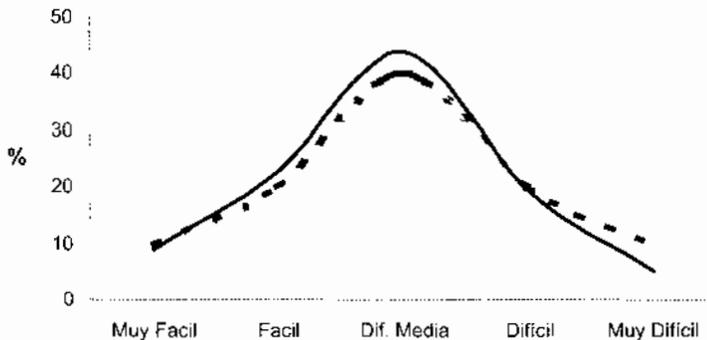
Tabla N° 6: Distribución del porcentaje empírico de los 57 ítems seleccionados en la matriz de especificaciones.

	Conceptos	Algoritmos	Problemas	Total
<b>Números/Operaciones</b>	10,5	28,1	21,0	59,6
<b>Geometría</b>	12,3	3,5	3,5	19,3
<b>Medidas</b>	7,1	3,5	1,7	12,3
<b>Organización Información</b>	1,7	3,5	3,5	8,7
<b>Total</b>	31,5	38,6	29,8	100

Si se comparan las dos tablas, se puede ver cómo los pesos empíricos de los ítems seleccionados replican de forma casi perfecta esperados teóricamente en la tabla de especificaciones. Hay algunos pequeños desfases. Así, en el eje de capacidades matemáticas hay menos conceptos de los esperados y más preguntas relativas a problemas. Por su parte, en el eje de los bloques de contenido, hay más preguntas referidas a números/operaciones y medidas y menos en el bloque de geometría. Sin embargo, estas desviaciones son en todos los casos tan pequeñas que es posible afirmar que los 57 ítems seleccionados respetan escrupulosamente la ponderación otorgada a cada entrada en la matriz de especificaciones original. Ello permite concluir que existen ciertas evidencias sobre la validez de contenido del banco de ítems.

Hay otro dato que habla de la bondad del proceso de selección de ítems. Es la distribución de los mismos en función de su índice de dificultad. Ya se vio en el gráfico 2 que los 117 reactivos seleccionados originalmente por su índice de discriminación se distribuían adecuadamente. En la selección de las 57 preguntas que irán al estudio final se ha procurado mantener esta misma proporción. Y a la vista del siguiente gráfico parece que se ha logrado.

Gráfico N.º 3: Comparación de la distribución teórica del porcentaje de ítems según su índice de dificultad y la distribución empírica de los 57 ítems seleccionados.



La línea punteada representa el porcentaje teórico de los ítems en función de su índice de dificultad. Es decir, 40% de ítems de dificultad media; 20% de ítems fáciles y difíciles; y 10 de ítems muy fáciles y muy difíciles. La línea sólida indica la distribución real conseguida. Los únicos desfases –por otro lado, muy pequeños– se deben a decisiones en la selección de ítems y no a la imposibilidad de replicar perfectamente la distribución teórica. En este caso se optó por incluir más ítems fáciles y de dificultad media y rebajar el número de ítems muy difíciles. La razón de esta decisión es doble. Por un lado, al considerar menos ítems muy difíciles se busca ampliar el máximo posible rango de los datos en el estudio final. Además, el hecho de incluir más ítems fáciles se debe a que las preguntas fueron pilotadas con alumnos que acababan de comenzar 3º de Educación Primaria y la prueba final se llevará a cabo con alumnos que van a finalizar el primer ciclo. Por ello, se espera que la inclusión de más ítems fáciles permitirá seguir manteniendo el índice de dificultad del banco de ítems en torno al 0,50 teórico.

De los 57 reactivos finalmente seleccionados 11 pertenecían al cuestionario A; 19 al B; 15 al C; y, 12 a la forma D. Con las pruebas reducidas únicamente a estos ítems se calcularon nuevamente los índices de fiabilidad (coeficiente alfa), puesto que serán los que se intenten mejorar en las aplicaciones finales. Como se puede ver en la tabla 7, el mejor comportamiento lo ofrece la Forma A, ya que, siendo un test muy corto –sólo 11 ítems–, se ha calculado un coeficiente alfa muy cercano a 0,80. En el extremo contrario se encuentra la forma D, que, como ya se venía viendo durante todo el estudio, es la escala menor. Su índice de fiabilidad, que apenas supera el 0,6, se antoja muy discreto. Por su parte, la forma B es la que ofrece una fiabilidad más alta y también interesante, puesto que supera el umbral del 0,8 aun siendo una prueba que no llega a 20 ítems. Por último, la forma C devuelve un índice moderado, aunque su longitud es también pequeña y con ítems similares podría estar en torno a 0,8 sólo con un pequeño incremento en su longitud. En definitiva el promedio de las cuatro pruebas es de 0,73, el cual se puede considerar aceptable para un primer pilotaje. Existen ciertas garantías de que en las pruebas finales, cuando el número de ítems dupliquen –en algunos casos– los de estas pruebas reducidas, se puedan conseguir índices alfa de Cronbach en torno a 0,85.

Tabla N° 7: *Índices de fiabilidad de las pruebas reducidas.*

	Nº ítems	Alfa de Cronbach
Forma A	11	0,78
Forma B	19	0,82
Forma C	15	0,71
Forma D	12	0,61
Promedio	57	0,73

Finalmente, en la tabla 8, se presentan los 57 ítems seleccionados. En la primera columna aparece la identificación del ítem (por ejemplo, C04 significa el ítem 4 del cuestionario C), a continuación hay una definición operativa de la pregunta. Las dos últimas columnas recogen los índices de dificultad y discriminación de cada ítem. Como se puede ver, la media de los índices de dificultad, expresada en porcentaje de aciertos, es de 52,86%, es decir muy cercano al 50% teórico. Por su parte, la media de los índices de discriminación de las preguntas es 0,449 que puede considerarse un resultado muy satisfactorio y que dice bastante de las posibilidades del banco de ítems para discriminar entre buenos y malos rendimientos.

## 6. Conclusiones

A la vista de los resultados presentados, la conclusión general es que la selección de ítems del estudio piloto se puede considerar satisfactoria. Se han logrado unos índices de fiabilidad interesantes y, en algunos casos, muy respetables a pesar de trabajar con test cortos –menos de 20 ítems–. Además, y atendiendo a los índices de dificultad de los reactivos, éstos se distribuyen normalmente y prácticamente no existe desviación de la proporción de aciertos teórica (0,5). Por otra parte, los ítems cubren todas las entradas de la matriz de contenidos con los pesos propuestos inicialmente, lo que permite ser positivos a la hora de enjuiciar el proceso de la selección muestral de los contenidos.

En definitiva, los datos hallados en el pilotaje *Mathematikoi-I* permiten esperar que, cuando estos ítems se prueben definitivamente en la aplicación final, se logren dos protocolos de evaluación fiables, discriminantes, ajustados a normalidad y válidos desde el punto de vista del contenido.

Tabla N.º 8: *Relación de los ítems seleccionados en el estudio piloto Mathematikoi-I con su porcentaje de aciertos y su índice de discriminación.*

	Dificultad	Discrimina.
C04 Conocer los lados de un triángulo	93,75	0,313
A10 Completar algoritmo de multiplicación (... x 3 = 12)	85,19	0,471
A25 Conocer la unidad de medida (o sistema de medida) adecuado para calcular pesos	84,38	0,668
D04 Formar un número dados centenas, decenas y unidades	83,87	0,543
A17 Problema sencillo de suma con números enteros	81,25	0,220
A22 Reloj analógico: identificar las 15:00	78,13	0,404
D02 Resta sencilla	77,42	0,539
D34 Problema complejo de varias sumas sucesivas	77,42	0,316
B19 Concepto básico: derecha	75,76	0,347
B03 Concepto de mitad (mitad de 20)	72,73	0,527
B08 Ordenar enteros de menor a mayor	72,73	0,357

D01	Algoritmo de suma con tres sumandos (2 de centenas y una de decenas)	70,97	0,390
B27	Desplazamientos básicas en el plano (3 izquierda, 1 arriba...)	69,70	0,249
C09	Concepto de decena	65,63	0,569
B34	Interpretar tabla horaria e identificar la hora de salida de un determinado tren	63,64	0,318
A23	Conocer los minutos que hay en media hora	62,50	0,641
D22	Litros en 4 medios litros	61,29	0,467
D28	Problema sencillo de división	61,29	0,528
C17	Resta con llevadas y cero en el medio (5042-3427)	59,38	0,362
D05	Algoritmo combinado de suma y resta	58,06	0,569
B16	Problema de suma y resta	57,58	0,501
B31	Suma de euros y céntimos	57,58	0,358
B35	Interpretar una tabla con varios criterios y contestar una pregunta sobre dos de ellas	57,58	0,261
A14	Concepto de multiplicación como suma rápida ( $8 \times 7 =$ sumar 7 veces 8)	56,25	0,546
C15	Identificar el eje de simetría correcto en un triángulo equilátero	56,25	0,513
C27	Problema dinero. Sumar céntimos, restar euros y expresar el resultado en céntimos	56,25	0,500
D29	Problema sencillo de resta	54,84	0,647
B05	Concepto de número impar (identificar uno entre varios pares)	54,55	0,721
B10	Completar algoritmo de resta (... - 9 - 5)	54,55	0,362
B28	Interpretar una escala de medida (distancia entre dos puntos)	54,55	0,125
B33	Interpretar un gráfico de barras y contestar una pregunta que implica resta sencilla	54,55	0,614
A21	Problema intermedio de suma y resta o de dos restas	53,13	0,445
A31	Identificar un cuerpo geométrico entre varios polígonos	53,13	0,448
C13	Completar serie descendente de enteros (2)	53,13	0,230
C32	Problema sencillo de multiplicación	50,00	0,228
B02	Escribir en letras números hasta el millar (728)	48,48	0,527
B32	Comparar centenas y unidades empleando los signos $>$ , $>$ , $=$	48,48	0,340
C05	Identificar una línea vertical	46,88	0,579
B04	Concepto de docena	45,45	0,554
B14	Problema sencillo de resta (concepto de faltar)	42,42	0,640
D06	Descomponer número en centenas, decenas y unidades	41,94	0,420
C07	Concepto de triple (triple de 20)	40,63	0,554
C30	Identificar el número de vértices que hay en un cubo	40,63	0,226
B11	Operación de multiplicación ( $367 \times 7$ )	39,39	0,296
D23	Sumar $3c$ , $2d$ y $4u + 3c$ y $3u$	38,71	0,408
D13	Dada tabla de valores energéticas calcular Kcal multiplicando por 10 y sumando	35,48	0,645
C24	Problema complejo de división y resta	31,25	0,419
C25	Dada tabla de valores energéticos calcular las Kcal mediante suma	31,25	0,363
C33	Saber cuántos metros hay en un kilómetro	31,25	0,328
D12	Interpretar un mapa de carreteras: distancia más corta entre dos puntos	29,03	0,697
A06	Completar una serie de enteros donde alternativamente se resta y se suma	28,13	0,421
A16	Comparar dos cantidades: una expresada en unidades y otra en decenas y unidades	28,13	0,391
C01	Identificar un pentágono	28,13	0,551
B15	Problema complejo de tres sumas combinadas	21,21	0,494
C22	Problema complejo de multiplicación y suma	18,75	0,417
A12	División simple (63:7)	9,38	0,768
B26	Calcular perímetro de un cuadrado	9,09	0,281
	<b>Media</b>	<b>52,86</b>	<b>0,449</b>

## Referencias bibliográficas

- Alonso Tapia, J. (dir.)(1997). *Evaluación del conocimiento y su adquisición*. Madrid: MEC/CIDE.
- Amón, J. (1984). *Estadística para psicólogos* (2 vols.). Madrid: Pirámide.
- Arnal, J. (1988). *Banco de ítems para la evaluación de las matemáticas en 5º de EGB*. Barcelona: PPU.
- Barbero, M.I. y Navas, M.J. (1995). *Creación de un sistema computerizado de evaluación de la capacidad matemática*. Madrid: CIDE/MEC (Memoria de investigación inédita).
- Bennett, G. y cols. (1985). *Test de Pronóstico Académico (APT)*. Madrid: TEA.
- Boehm, A.F. (1986). *Test Boehm de Conceptos Básicos*. Madrid: TEA.
- Coll, C. y Miras, M (1990). *Características individuales y condiciones de aprendizaje: la búsqueda de interacciones*. En Coll, Palacios y Marchesi (Comp.): *Desarrollo psicológico y educación*. Vol II, págs. 395-417. Madrid: Alianza.
- EOEP-Nalón (2001). Descripción de una prueba piloto para medir el nivel de competencia curricular en el área de matemáticas al final de la educación primaria. *Boletín de la ASPOPA*, 6, 5-11.
- Fernández Alonso, R. y otros (2003). *Mathematikoi-III. Manual técnico*. Oviedo: Consejería de Educación y Cultura del Principado de Asturias.
- González Suárez, M.D. (1989). *Exploración de las dificultades individuales de lectura (EDIL)*. Madrid: TEA.
- INCE (1996). *Lo que aprenden los alumnos de 12 años. Evaluación de la educación primaria. Datos básicos 1995*. Madrid: INCE/MEC.
- INCE (1997). *Evaluación de la educación primaria*. Madrid: INCE/MEC.
- INCE (1998). Los resultados escolares. En *Diagnóstico del sistema educativo. La escuela secundaria obligatoria. 1997*. Madrid: INCE/MEC.
- INCE (2000). *Evaluación de la educación primaria. Datos básicos 1999*. Madrid: INCE/MEC.
- Lapointe, A.E., Mead, N.A. y Phillips, G.W. (1989). *Un mundo de diferencias. Un estudio internacional de evaluación de las matemáticas y las ciencias*. Madrid: MEC/CIDE.
- MEC (1992). *Materiales para la reforma*. Madrid: MEC.
- Pérez Avellaneda y otros (1999). *Batería de contenidos escolares de primaria (BACEP)*. Madrid: CEPE.
- Sociedad Andaluza de Educación Matemática Thales (1991). *Estándares curriculares y de evaluación para la educación matemática*. Sevilla: SAEM.
- Thurstone, L.L. y Thurstone, TH.G. (1979). *Test de Aptitudes Escolares (TEA)*. Madrid: TEA.
- Yuste, C. (1985). *Batería de Aptitudes Diferenciales y Generales (BADYG)*. Madrid: CEPE.
- Yuste, C. (1995). *Inteligencia General y Factorial (IGF)*. Madrid: TEA.
- Yuste, C. y Sanz, L. (1991). *Conceptos Básicos Relacionantes (COBAR)*. Madrid: CEPE.