APLICACIÓN Y VALIDACIÓN DE UN PROCEDIMIENTO DE CONSTRUCCIÓN DE PRUEBAS DE RENDIMIENTO DE MATEMÁTICAS, CIENCIAS Y LENGUA EN LA EDUCACIÓN PRIMARIA

GUILLERMO GIL ESCUDERO
JUAN CARLOS SUÁREZ FALCÓN (*)

INTRODUCCIÓN

La Ley Orgánica de Ordenación General del Sistema Educativo (LOGSE) creó, en su artículo 62, el Instituto Nacional de Calidad y Evaluación (INCE) encargándole al mismo, entre otras, la tarea de llevar a cabo la evaluación general del sistema educativo en un marco de respeto a la distribución de competencias entre Comunidades Autónomas, tal y como aparece regulado en la Constitución Española y en los Estatutos de Autonomía. Además de las actividades de innovación e investigación educativa, la orientación educativa y profesional, la inspección técnica de educación y la mejora y reorientación de aspectos tales como la formación del profesorado, las programaciones docentes y la función directiva, la evaluación nacional del sistema educativo en España es también una medida que contribuye a la mejora de la calidad de la enseñanza.

Uno de los objetivos del INCE consiste en obtener indicadores del funcionamiento del sistema educativo en su conjunto que permitan conocer el grado en que el sistema educativo español logra alcanzar, en un momento determinado de su desarrollo, los objetivos establecidos por las leyes para la educación. La finalidad básica de este objetivo es la de proporcionar información relevante no sólo a las Administraciones Educativas, sino también a los miembros de la comunidad educativa y a los ciudadanos en general. Por otro lado, el INCE debe también analizar qué factores influyen sobre los procesos escolares, en especial los procesos de enseñanza y aprendizaje, de modo que dichos análisis permitan aplicar medidas que contribuyan a mejorar la calidad de la educación.

El INCE, a lo largo del año 1995, llevó a cabo su primer proyecto de evaluación, que consistió en la evaluación de la Educación Primaria. En este proyecto se pretendía analizar los resultados educativos alcanzados al final de esta etapa educativa, esto es, al término de sexto curso de Educación Primaria. Como este nivel, en el momento de llevar a cabo el proyecto, no se había implantado en la totalidad del Estado, se recopiló información sobre los resultados obtenidos por el alumnado al finalizar el sexto curso de Enseñanza General Básica (EGB), como un primer paso

(*) Instituto Nacional de Calidad y Evaluación.

que sirviera de línea base para poder llevar a cabo un estudio comparativo que ha de desarrollarse cíclicamente en años futuros.

El estudio establece además, un punto de referencia para evaluar, a lo largo del tiempo, los resultados educativos proporcionados por el conjunto del sistema, así como los efectos de los cambios producidos por la nueva ordenación educativa y por las diferentes medidas de gobierno de la educación sobre aspectos tales como el desarrollo global de los alumnos, el rendimiento académico, el desarrollo de las actitudes, el funcionamiento de los centros, la participación de los miembros de la comunidad educativa, etc.

En cuanto al rendimiento académico, objeto de este trabajo, se estudió el grado de adquisición de las enseñanzas mínimas al final de la Educación Primaria en las áreas de Lengua Castellana, Matemáticas, Conocimiento del Medio (Ciencias Naturales y Ciencias Sociales) y Educación Física.

La construcción de unas pruebas de rendimiento válidas y fiables se ha abordado desde diferentes orientaciones en el ámbito de la investigación y la evaluación educativas. Se han tenido en cuenta tanto la tradicional Teoría Clásica de los Tests (TCT) que actualmente sigue manteniendo una fuerte vigencia, como la Teoría de la Respuesta al Ítem (TRI) que se está considerando, cada vez con más fuerza, como una forma válida y fiable de abordar las tareas de medición de los resultados educativos.

Este trabajo pretende mostrar un procedimiento de construcción de pruebas de rendimiento con respuestas cerradas de opción múltiple, destinadas a evaluar el grado en el que los alumnos alcanzan los objetivos educativos definidos en el Real Decreto 1006/1991 de 14 de junio por el que se establecen las enseñanzas mínimas para la Educación Primaria en las áreas curriculares antes mencionadas.

El procedimiento propuesto para la construcción de las pruebas se fundamenta en un método mixto que trata de conjugar, por una parte, la evaluación criterial con la normativa y, por otra, los métodos de la TCT con los de la TRI. No obstante, la aplicación de estos métodos tiene ciertas restricciones que es necesario mencionar para entender el procedimiento seguido en la construcción de las pruebas. La primera restricción viene dada por el propio cometido del INCE, que consiste en comprobar en qué grado los alumnos de sexto curso alcanzan las enseñanzas mínimas, lo que lleva a tener cuenta, en primer lugar, que la selección de los elementos de evaluación no es totalmente libre, sino que está determinada en gran medida, por el marco curricular prescrito en las enseñanzas mínimas. Por tanto, a la hora de seleccionar los ítems no se trata simplemente de seleccionar aquellos que funcionan mejor desde un punto de vista psicométrico, sino que se han de seleccionar también en función de las categorías establecidas por dicho marco curricular, es decir, en función de los bloques temáticos y los contenidos asociados a los mismos, de manera que no quede excluido ninguno de los bloques temáticos que conforman cada una de las áreas. La segunda restricción viene dada por el hecho de que al tratarse de la evaluación de las enseñanzas mínimas, los ítems deben corresponderse con los contenidos que teóricamente deberían enseñarse a todos los alumnos.

Ahora bien, estas restricciones dejarían de considerarse como tales si se las considerara como requisitos a tener en cuenta en el aspecto criterial de esta evaluación, teniendo en cuenta que, tal y como señalan Hambleton y Rogers (1989), para muchos propósitos educativos, la determinación del nivel de habilidad de los examinados es substancialmente más importante que la determinación de la situación del examinado en relación con un grupo normativo, por lo que en estos casos, el marco de evaluación referido a una norma es menos apropiado.

Sin embargo, como ya se ha mencionado, también es cometido del INCE investigar y analizar los factores que influyen sobre los resultados educativos, especialmente aquellos que sean susceptibles de modificación por parte de los agentes educativos (administraciones, profesores, familias, etc.), para que puedan proporcionar una mejora de la calidad educativa. El hecho de tener en cuenta estos factores proporciona a la evaluación un carácter normativo.

Esta doble finalidad que deben cumplir las evaluaciones competencia del INCE, origina la necesidad de hacer compatibles los dos tipos de evaluación (la criterial y la normativa) dentro de un mismo estudio. Este doble análisis de una misma realidad pone de manifiesto las dificultades técnicas que surgen a la hora de hacer coexistir dos metodologías diferentes, tanto en sus orígenes como en sus procedimientos de desarrollo y aplicación. Por ello, ha sido necesario idear un procedimiento de construcción de pruebas de rendimiento que reduzca al mínimo las posibles incompatibilidades entre una y otra metodología.

PROCEDIMIENTO

Prueba piloto

Una vez finalizado el diseño del proyecto, la primera fase de trabajo consistió en un análisis exhaustivo de la legislación educativa vigente (Constitución, LODE, LOGSE, Real Decreto 1006/1991). A partir de este análisis se pudieron definir unas matrices conceptuales o especificaciones que serían la base, tanto de cada una de las pruebas de las áreas que iban a ser evaluadas, como de los cuestionarios de opinión con los que se iba a recabar información de los distintos agentes que forman parte del ámbito educativo. Las matrices de las pruebas de rendimiento de las distintas áreas estaban formadas por los bloques temáticos con sus correspondientes conceptos, procedimientos y actitudes, a partir de los cuales se establecían las subáreas y los elementos de evaluación. Estas matrices especificaban tanto el tipo de contenidos que deberían examinarse, como el peso que cada uno de los subapartados debía tener en el conjunto de las pruebas.

Las Tablas I, II y III presentan las matrices conceptuales para las áreas de Lengua Castellana y Literatura, Matemáticas y Conocimiento del Medio (Ciencias Naturales y Ciencias Sociales) indicando el peso relativo de cada uno de los contenidos y de las capacidades en el conjunto de cada una de las pruebas. Establecidas estas especificaciones para la construcción de las pruebas, quince profesores especialistas en este nivel educativo y en ejercicio, asesorados por los expertos en evaluación del INCE, redactaron aproximadamente 600 ítems por materia que se correspondían con los contenidos determinados por las matrices conceptuales del diseño. Posteriormente, se seleccionaron aproximadamente 200 ítems por área. Con ellos se formaron las cuatro pruebas que se evaluaron en el estudio piloto. Cada una de estas pruebas no sólo tenía un número de ítems similar, sino que también era similar su construcción y el nivel de su dificultad.

Estas pruebas fueron también revisadas por los expertos designados por las Comunidades Autónomas con competencias transferidas en educación y por el MEC para el resto de las comunidades. Las diversas sugerencias y correcciones que hicieron se incorporaron al proceso de construcción de las pruebas.

El muestreo se llevó a cabo mediante un procedimiento estratificado y aleatorio, formándose los estratos en función: *a*) del nivel de fracaso escolar del centro (alto, medio y bajo) según la información aportada por las Administraciones Educativas; *b*) de la lengua propia de la comunidad autónoma y *c*) de la titularidad del centro (público o privado). Teniendo en cuenta estos estratos, se seleccionaron 35 centros en los que se probaron las cuatro pruebas elaboradas para cada área.

TABLA I

Distribución de los ítems de las cuatro pruebas del área de Lengua Castellana del estudio piloto, según los ítems de lectura y escritura y los tipos de textos

LECTURA		
Tipo de textos		Total
Literarios	Descriptivos, Diálogos, Monólogos, Narrativos, Poéticos, Teatrales, etc.	49
Informativos	Recetas, Instrucciones, Divulgativos, Periodísticos, Cartas, Avisos, etc.	42
Verbal y no verbal	Comics, Viñetas, Anuncios gráficos, Jeroglíficos, Ilustraciones, Planos, etc.	24
Total		115
ESCRITURA		
	Ortografía, Puntuación, Vocabulario, Sinónimos, Ordenación de palabras y frases, Continuación del escrito, Incorrecciones gramaticales, Formación de oraciones, Estilos directo e indirecto, etc.	52
TOTAL		167

TABLA II

Distribución de los ítems de las cuatro pruebas del área de Matemáticas del estudio piloto, según los contenidos y las capacidades evaluables

CONTENIDOS		CAPACIDADES			
		Conocimiento conceptual	Procedimientos Estrategias	Problemas	Total
Números y Operaciones	Números naturales Valor posicional	4	2	0	6
	Divisibilidad	4	4	10	18
	Operaciones Propiedades	4	12	7	23
	Potencias Raíces cuadradas	4	2	3	9
	Números quebrados	4	8	6	18
	Números decimales	4	8	4	16
	Otros temas	4	4	2	10
Medida de magnitudes		8	12	12	32
Geometría	Elementos geométricos del plano	12	6	4	22
	Elementos geométricos del espacio	8	1	1	10
	Perímetros, Áreas, Volúmenes	4	1	15	20
Organización de la información	Representación y análisis de datos Probabilidad	4	4	0	8
Total		64	64	64	192

Conocimiento conceptual: Capacidad para recordar conceptos matemáticos.
Procedimientos estrategias: Capacidad para recordar y utilizar procedimientos y estrategias.
Problemas: Capacidad para solucionar problemas utilizando el conocimiento matemático.

TABLA III

Distribución de los ítems de las cuatro pruebas del área de Conocimiento del Medio (Ciencias Naturales y Ciencias Sociales) del estudio piloto, según los contenidos y las capacidades evaluables

CONTENIDOS	CAPACIDADES				Total
	Conocimiento	Comprensión	Aplicación	Análisis	
El ser humano y la salud	4	8	8	4	24
Paisaje	6	9	9	6	30
Medio físico	4	8	8	4	24
Seres vivos	4	6	6	4	20
Materiales y sus propiedades	4	6	6	4	20
Población y actividades humanas	4	6	6	4	20
Máquinas y aparatos	2	4	4	2	12
Organización social	2	3	3	2	10
Medios de comunicación y transporte	2	3	3	2	10
Cambios y paisajes históricos	6	8	9	7	30
Total	38	61	62	39	200

Conocimiento: Capacidad para recordar conceptos.
Comprensión: Capacidad para explicar e interpretar información.
Aplicación: Capacidad para aplicar los conocimientos a nuevas situaciones y a la solución de problemas.
Análisis: Capacidad para analizar y valorar informaciones y datos.

La aplicación de las pruebas la llevó a cabo una empresa especializada contratada al efecto, supervisada por observadores del INCE. Cada tipo de prueba fue contestada por aproximadamente 400 alumnos de sexto de EGB. También a 400 alumnos que respondieron a uno de los modelos de las pruebas se les administró un cuestionario de opinión. Adicionalmente, se probaron en la prueba piloto los cuestionarios de profesores, de los equipos directivos y de las familias de los alumnos.

PROCEDIMIENTO DE CONSTRUCCIÓN DE LAS PRUEBAS

Los manuales básicos, tales como los de Bejar (1983), Linn (1989), Mehrens y Lehman (1984), Popham (1978, 1980, 1993), Walberg y Haertel (1990), exponen los diferentes métodos de construcción de tests y en especial los Test Referidos a un Criterio (TRCs). Hambleton (1980, 1982) y Popham (1978, 1980) han centrado su trabajo en la revisión de los métodos de preparación y validación de conjuntos de ítems referidos a un criterio. Navas (1994) hace una revisión en castellano de algunas aportaciones y desarrollos de la construcción de los TRCs en la que destaca varias opiniones que señalan la falta de adecuación de los métodos de la TCT para la selección de ítems (método aleatorio y método clásico) en los TRCs. Sin embargo, en esta misma revisión se señala que la TRI parece constituir una aproximación que puede contribuir a una mejora significativa de los procedimientos de construcción de los TRCs.

Desde esta perspectiva, se distinguen fundamentalmente dos métodos: *a)* el método óptimo de selección de ítems, que establece una puntuación de corte como criterio y selecciona los ítems de modo que su discriminación sea máxima en el punto del rasgo a evaluar asociado a la puntuación de corte y, dentro de éstos, selecciona los ítems que proporcionan mayor información en ese punto de la escala; y *b)* el método óptimo de contenido, que es similar al anterior, pero que, además, establece una restricción al test final que consiste en que los contenidos deben satisfacer una serie de condiciones, ya que han de ajustarse a un conjunto predeterminado de los mismos, establecidos por las matrices conceptuales que previamente han sido diseñadas para cada área que va a ser objeto de evaluación. Los trabajos de Hambleton y Rogers (1988), de Gruijter y Hambleton (1982) y Hambleton y de Gruijter (1983), ponen de manifiesto las ventajas y la mejor adecuación de estos dos últimos métodos sobre los basados exclusivamente en la TCT.

En la selección de ítems que se llevó a cabo se tuvieron en cuenta las dos perspectivas: La Teoría Clásica de los Tests (TCT) y la Teoría de la Respuesta al Ítem (TRI) utilizando un modelo de tres parámetros (ítems de elección múltiple). En primer lugar, se comprobó la unidimensionalidad de las pruebas tras realizar un análisis factorial para variables dicotómicas (Bock y Aitkin, 1981) mediante el programa TESTFACT (Wilson, Wood, Kandola y Gibbons, 1991) en cada una de las cuatro pruebas piloto de cada área curricular. En las dieciséis pruebas se verificó la existencia de un único factor significativo por prueba, lo que permitía considerar a cada una de ellas como una prueba unidimensional.

El primer criterio para la selección de ítems procedió de la TRI. Se eliminaron todos aquellos ítems cuyas probabilidades asociadas a χ^2 fueran inferiores a .10 bajo el modelo logístico de tres parámetros, obtenidas con el programa BILOG.3 (Mislevy, R; y Bock, D., 1990). A continuación se utilizaron varios criterios de selección complementarios basados en la TCT. Se fijó una discriminación mínima del ítem (.20), tomando como índice de discriminación la correlación biserial puntual (calculada con el programa TESTFACT).

Por otra parte, se analizaron el porcentaje de respuestas dadas por los alumnos a los distractores y se tuvo en cuenta que dichas respuestas estuvieran adecuadamente distribuidas, es decir, que los distractores actuaran como tales, de modo que si el porcentaje de elección de algunos de los distractores de un ítem era nulo o desequilibrado y pedagógicamente inexplicable, no se seleccionaba dicho ítem.

Todos los ítems que cumplieron estos criterios se distribuyeron por grupos en función de las matrices conceptuales del diseño y se eligieron los mejores ítems de cada apartado de las matrices, esto es, el mejor o los mejores ítems para cubrir cada objetivo mínimo especificado.

Una vez seleccionados los ítems en función de su calidad psicométrica y su ubicación en las matrices, otro problema que hubo que resolver para elaborar las pruebas finales de rendimiento, consistió, como ya se mencionó anteriormente, en hacer converger las dos formas diferentes de abordar la construcción de los instrumentos de evaluación: la criterial y la normativa. Si se consideraba exclusivamente el nivel óptimo en los resultados de la evaluación criterial, cabría esperar curvas de rendimiento asimétricas negativas y es conocido que este tipo de distribuciones viola algunos supuestos básicos a la hora de realizar análisis estadísticos posteriores, como es el supuesto de normalidad de la distribución en las pruebas paramétricas asociadas a la evaluación normativa.

En consecuencia, se diseñó un mecanismo que, sin distorsionar los resultados originales, respetase la distribución de facilidad/dificultad para cada área que fue obtenida de la totalidad de los ítems de las

cuatro pruebas del estudio piloto ya que, como puede suponerse, los mejores ítems procedían de pruebas diferentes. Los porcentajes que marcaban el nivel de dificultad de los ítems se clasificaron en cinco grupos, de menor a mayor dificultad, con el fin de que estuvieran representados en la prueba final de la forma más semejante posible. De este modo, se seleccionaron los ítems que conformaron las pruebas definitivas. El orden en que aparecen los ítems es el de su progresiva dificultad (excepto en Lengua Castellana y Literatura, porque distintos grupos de ítems están asociados a distintos tipos de texto, y por ello, lo que se ordenó en esta prueba fueron los textos de menor a mayor dificultad media de los ítems asociados a los mismos).

ADMINISTRACIÓN DE LAS PRUEBAS FINALES

En mayo de 1995 se aplicaron las pruebas definitivas de Lengua Castellana y Literatura, Matemáticas y Ciencias Sociales y Ciencias Naturales a 10.870 alumnos de sexto curso de la EGB, junto con los cuestionarios dirigidos a los padres, los equipos directivos, y a los propios alumnos. Como en el caso de la prueba piloto, de la administración de las pruebas se encargó una empresa especializada cuyo trabajo fue dirigido y supervisado por técnicos del INCE.

El tamaño de la muestra se determinó a partir de los criterios establecidos por la *International Association for the Evaluation of Educational Achievement* (IEA) (Ross, 1991) de forma que fuese representativa a nivel del Estado. Si se siguen las tablas del diseño de muestras para muestreos bietápicos construidas por Ross (1987) y se considera un coeficiente de correlación intraclase (ρ) igual a .3, con una ratio mínima estimada de 20 alumnos por clase, se comprueba que es necesario, para que los resultados obtenidos a partir de los alumnos sean representativos del conjunto del Estado, muestrear un mínimo de 134 clases en todo el Territorio Nacional, de manera que los cálculos estadísticos se muevan en un margen de error estimado de $\pm$ 0.1σ para las medias, $\pm$ 5 por 100 para los porcentajes y $\pm$ 0.1 para los coeficientes de correlación.

Las variables que se tuvieron en cuenta para realizar la estratificación fueron la comunidad autónoma, titularidad del centro –pues parece conveniente conocer de forma diferenciada el funcionamiento de los dos tipos de centros: públicos y privados– y el tamaño de los mismos, ya que éste implica diferencias organizativas internas.

El muestreo que se realizó fue proporcional en lo que se refiere a las categorías de titularidad y tamaño de los centros, pero no lo fue respecto de los estratos por comunidades autónomas. Esta no proporcionalidad por comunidades se hizo con el objeto de poder establecer comparaciones entre las diferentes comunidades autónomas con competencias transferidas en educación (Andalucía, Canarias, Cataluña, Galicia, Navarra, País Vasco y Valencia) y entre éstas y el territorio gestionado provisionalmente por el Ministerio de Educación.

Para realizar estas comparaciones se fijó un margen inicial de error estimado de $\pm$ 0.2σ para las medias y $\pm$ 10 por 100 para los porcentajes, lo que implicaba muestrear un conjunto de 34 centros por comunidad autónoma con competencias transferidas en educación. Los excesivos costos de aplicación llevaron a tomar la determinación de que en el territorio gestionado por el MEC se muestrearan 20 centros por comunidad autónoma (Aragón, Asturias, Baleares, Cantabria, Castilla-La Mancha, Castilla-León, Extremadura, La Rioja, Madrid y Murcia) y no los 34 como en el resto de las comunidades, lo que implicaba un conjunto de 200 clases y unos márgenes de error algo mayores que para el resto de comunidades autónomas. No obstante, hay que tener en cuenta que las

estimaciones realizadas para marcar los márgenes de error son conservadoras y, además, el coeficiente de correlación intraclase es esperable que sea menor al disminuir el ámbito de su estimación del conjunto del Estado al de las comunidades autónomas, con lo que se espera que haya una reducción de la variabilidad entre clases.

La determinación de los centros específicos que participaron en la evaluación se realizó por muestreo aleatorio, estratificado y con probabilidad proporcional al tamaño. Dentro del centro, la elección de un grupo-clase concreto del curso, en el caso de que hubiera más de uno, se realizó por muestreo aleatorio.

VALIDACIÓN DEL PROCEDIMIENTO DE SELECCIÓN DE ÍTEMS

Se han utilizado tres criterios para la validación del procedimiento de selección de ítems. El primer criterio consistió en fijar para la prueba final una fiabilidad superior a los índices de fiabilidad obtenidos en cada una de las cuatro pruebas que para cada área se hicieron en el estudio piloto. El segundo criterio se basó en que los índices de facilidad y discriminación de los ítems no variaran de forma sustancial de una aplicación a otra. Por último, el tercer criterio consistió en que los ítems de la prueba final se ajustasen, en su gran mayoría, al modelo logístico de tres parámetros.

La Tabla IV ofrece los resultados relativos al primer criterio sobre los índices de fiabilidad de las cuatro pruebas piloto y de la prueba final de las diferentes materias. Como puede apreciarse, la fiabilidad de la prueba final supera, en todos los casos, tanto la máxima fiabilidad lograda en cada una de las cuatro pruebas del estudio piloto para cada una de las áreas evaluadas, como su promedio. El mayor incremento en fiabilidad se alcanzó en la prueba de Lengua Castellana y Literatura, mientras que en la prueba de Matemáticas, sólo se obtuvo una mínima ganancia en precisión con respecto al estudio piloto. En cualquier caso, el procedimiento de selección de ítems utilizado, permite disponer de unas pruebas que tienen una alta fiabilidad para evaluar el rendimiento de los alumnos en las áreas analizadas.

TABLA IV

Índices de fiabilidad de las cuatro formas de las pruebas piloto y las pruebas finales en Matemáticas, Lengua y Conocimiento del Medio (Ciencias Naturales y Ciencias Sociales)

Materia	Prueba piloto					Prueba final
	Forma A	Forma B	Forma C	Forma D	Promedio	
Matemáticas	0,822	0,841	0,837	0,843	0,835	0,854
Lengua	0,815	0,770	0,814	0,810	0,802	0,894
Ciencias	0,808	0,767	0,849	0,833	0,814	0,872

TABLA V

Correlaciones entre los índices de facilidad y discriminación de los ítems comunes en la aplicación del estudio piloto y en estudio final por áreas de conocimiento

Materia	Índices de facilidad	Índices de discriminación
Matemáticas	0,9456 n = 40 p = 0,000	0,69 n = 40 p = 0,000
Lengua	0,9527 n = 63 p = 0,000	0,61 n = 63 p = 0,000
Ciencias	0,9531 n = 50 p = 0,000	0,72 n = 50 p = 0,000

Para la verificación del segundo criterio, se analizaron las correlaciones entre las estimaciones paramétricas de los ítems comunes de las pruebas del estudio piloto y de la prueba final como técnica para comprobar su grado de variación. Una correlación positiva y alta supone que la variación es mínima y no significativa. En la tabla 5 se presentan las correlaciones entre los índices de discriminación y facilidad de los ítems de las dos aplicaciones en las áreas evaluadas. Se observa que existe una relación muy fuerte entre los índices de facilidad de los ítems ya que entre ellos hay una correlación de 0,95 en cada una de las áreas evaluadas.

En el caso de la discriminación, la relación lineal es elevada (media de 0,67) aunque no tan alta como en las estimaciones de la facilidad de los ítems. Un análisis global de estas relaciones, indica que existe una estructura correlacional entre las estimaciones paramétricas que permanece con independencia del área analizada.

Por último, en lo que concierne al ajuste de los ítems a los modelos logísticos, hay que comentar que las tres pruebas finales se ajustan adecuadamente al modelo logístico de tres parámetros (nivel de significación $\alpha = 0,05$), ya que sólo se desajusta un ítem en Matemáticas, dos en Lengua Castellana y Literatura y tres ítems en la prueba de Ciencias Sociales y Ciencias Naturales, lo que supone que más del 90 por 100 de los ítems en cada prueba se ajusta a este modelo.

COMENTARIOS FINALES

En el trabajo que aquí se presenta se ha propuesto un procedimiento mixto para la selección de ítems y construcción de pruebas de rendimiento que intenta combinar, por un lado, la evaluación criterial con la normativa y, por otro, los métodos de la Teoría Clásica de los Tests (TCT) con los de la Teoría de la Respuesta al Ítem (TRI).

La utilización de este procedimiento para la construcción de tres pruebas de rendimiento en Lengua Castellana y Literatura, Matemáticas, y Ciencias Sociales y Ciencias Naturales de sexto curso de EGB, ha proporcionado resultados que indican que las características psicométricas de los ítems seleccionados, tales como los índices de discriminación y facilidad y el ajuste al modelo logístico de tres parámetros, se preservan en la prueba final, en tanto que la precisión o fiabilidad global de la prueba definitiva supera ligeramente la fiabilidad esperada, si se toma como punto de

partida las cuatro formas de las pruebas del estudio piloto.

En conclusión, los resultados obtenidos parecen indicar que el procedimiento mixto utilizado es un mecanismo válido para la construcción de pruebas de rendimiento en las que la selección de ítems no puede realizarse atendiendo únicamente al funcionamiento psicométrico de esos ítems, sino que dicha selección ha de tener en cuenta, en este caso, las categorías establecidas por el marco curricular para las materias evaluadas.

Por otro lado, como este procedimiento considera criterios derivados de las dos grandes aproximaciones teóricas de la psicometría, la TCT y la TRI, en él se combinan las ventajas de la TCT (generalidad de aplicación, supuestos débiles o pocos restrictivos, mayor manejabilidad, menor sofisticación matemática, etc.) con las de la TRI (invariación de los parámetros de los ítems y de la aptitud, falsabilidad de los modelos, medidas locales de precisión, etc.)

Con el objetivo de replicar estos resultados, está previsto aplicar el procedimiento mixto antes descrito a otras áreas y niveles educativos. En principio, se esperan resultados similares a los encontrados en la construcción de las tres pruebas de rendimiento de sexto curso.

Estos estudios proporcionarán una indicación del grado de generalización por niveles y áreas curriculares, así como una validación global del procedimiento propuesto para la construcción de pruebas de rendimiento para la evaluación de la enseñanza no universitaria.

BIBLIOGRAFÍA

BEJAR, I. I.: *Achievement Testing: Recent Advances*. Beverly Hills, Sage, 1983.

BERK, R. A.: «Criterion Referenced Tests», en WALBERG, H. J. y HAERTEL, G.D. (eds.), *The International Encyclopedia of Educational Evaluation*, Oxford, Pergamon, 1990.

GRUIJTER, D. N. M. de y HAMBLETON, R. K.: «Using item response models in criterion referenced test item selection», en HAMBLETON, R. K. (ed.), *Applications of item response theory*, Vancouver, BC: Educational Research Institute of British Columbia, 1983, 20, 4, pp. 355-367.

HAERTEL, G. D.: «Achievement Tests», en WALBERG, H. J. y HAERTEL, G.D. (eds.), *The International Encyclopedia of Educational Evaluation*. Oxford, Pergamon, 1990.

HAMBLETON, R. K.: «Test score validity and standard-setting methods», en BERK, R. (ed.), *Criterion-Referenced measurement: State of the art*, Baltimore: Johns Hopkins University Press, 1980.

– «Advances in criterior-referenced testing technology», en REYNOLDS, C. & GUTKIN, T. (eds.), *Handbook of school psychology*, New York: John Wiley & Sons, 1982.

HAMBLETON, R. K. y GRUIJTER, D. N. M. de: «Applications of item response models to criterion-referenced test item selection», *Journal of Educational Measurement*, 1983, 20, 4, pp. 355-367.

HAMBLETON, R. K. y ROGERS, H. J.: «Solving criterion-referenced measurement problems with item response models», *International Journal of Educational Research*, 1989, 13, 2, pp. 145-160.

LINN, R. L.: *Educational Measurement*, New York, Macmillan, 1989.

MEHRENS, W. A. y LEHMAN, I. J.: *Measurement and Evaluation in Education and Psychology*, New York, Holt, 1984.

MISLEVY, R. J. y DARRELL BOCK, R.: *Bilog 3. Item Analysis and Test Scoring with Binary Logistic Models*, Scientific Software, Inc., Mooresville, 1990.

NAVAS, M. J.: «Teoría Clásica de los Tests versus Teoría de Respuesta al Ítem», *Psicológica* 15, 1994, pp. 175-208.

POPHAM, W. J.: *Criterion-referenced measurement*, Englewood Cliffs, N. J.; Prentice Hall, 1978.
- *Modern educational measurement*, Englewood Cliffs, N. J.; Prentice Hall, 1981.
- *Educational Evaluation*, Boston, Allyn and Bacon, 1993.

REAL DECRETO 1006/1991, de 14 de junio, por el que se establecen las enseñanzas mínimas correspondientes a la Educación Primaria. BOE número 152, de 26 de junio de 1991.

ROSS, K. N.: «Sample Design», *International Journal of Educational Research*, 1987, 11, pp. 1-143.
— *Sampling Manual for the IEA International Study of Reading Literacy*, International Coordinating Center: IEA International Study of Reading Literacy: University of Hamburg, 1991.

WILSON, D. T.; WOOD, R.; KANDOIA, P. y GIBBONS, R.: *Testfact. Test scoring. Item Statistics, and Item Factor Analysis*, Scientific Software, Inc., Chicago, 1991.