Análisis de la dimensionalidad en modelos de valor añadido: estudio de las pruebas de matemáticas empleando métodos no paramétricos basados en TRI

Analysis of dimensionality at value-added models: a study of math test using non parametric methods based on item response theory

Luis Lizasoain Hernández
Luis Joaristi Olariaga
Universidad del País Vasco-Euskal Herriko Unibertsitatea. Departamento de Métodos de Investigación y Diagnóstico en Educación (MIDE). Donostia – San Sebastián, España.

Resumen

La evaluación basada en el valor añadido implica una metodología en la que se asumen supuestos fuertes. Se hace uso de diseños longitudinales y las puntuaciones de cada una de las mediciones deben ser situadas en una escala común. Ese escalamiento común supone un sólido diseño de equiparación. La única forma para asegurar la posibilidad de comparación dentro de los diseños longitudinales es emplear un modelo de medición donde toda escala comparte las mismas propiedades métricas. La posibilidad de la estimación en una escala común se basa en el supuesto de independencia local. El objetivo de este trabajo es analizar la estructura dimensional de las pruebas empleadas para evaluar el rendimiento académico en la asignatura de matemáticas en el contexto de la evaluación realizada en la Comunidad de Madrid en los cursos académicos 2005-06 y 2006-07 en tres cohortes: 5º y 6º de Educación Primaria, 1º-2º y 3º-4º de ESO. Para ello se evalúa su posible «unidimensionalidad» y la simplicidad o complejidad de su estructura empleando métodos no paramétricos basados en la Teoría de la Respuesta Ítem (TRI). Los resultados confirman estructuras esencialmente unidimensionales.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

A su vez, se comprueba que el aumento de la complejidad de los contenidos implica un aumento en la complejidad de la estructura dimensional de las pruebas. Los modelos de medida utilizados son ciertamente bastante robustos frente a las violaciones leves de los supuestos relacionados con la independencia local.

Palabras clave: dimensionalidad, DIMTEST, DETECT, escalamiento vertical, valor añadido.

Abstract

The added value based assessment implies the use of a methodology where strong assumptions are assumed. On one hand, longitudinal designs are used and, on the other hand, the scores of every time-moment have to be recoded into a common scale. This common scaling implies a robust equating design. The only way to ensure the comparability within the longitudinal designs is to use a measurement model where every scale shares the same metric properties. The possibility of common scale estimates is based upon the local independence assumption. The aim of this paper is to analyze the dimensional structure of a set of mathematics achievement tests used to assess the academic achievement in the Community of Madrid during 2005-06 and 2006-07 years and in three cohorts corresponding to the following academic degrees: 5th-6th of Primary Education, 1st-2nd and 3-4th of Obligatory Secondary Education. Their essential unidimensionality and the level of simplicity/complexity of their structure are assessed using nonparametric IRT-based procedures. These results confirm that the most of the tests have an essential unidimensional structure. In turn, the more of the contents complexity, the more of the tests dimensional structure complexity. The measurement models that have been used are certainly quite robust to minor violations related to the local independence assumptions.

Key Words: Dimensionality, DIMTEST, DETECT, vertical scaling, added value.

Introducción

La evaluación basada en el valor añadido implica una metodología en la que se asumen algunos supuestos fuertes. Por un lado, se hace uso de diseños longitudinales en los que cada individuo es medido en varias ocasiones, y por otro, estas puntuaciones deben ser situadas en una escala común a todas ellas. Ese escalamiento común supone un sólido diseño de equiparación. Nos encontramos así con dos restricciones concurrentes. Las pruebas que se usan en cada ocasión deben tener validez curricular, es

decir, su contenido debe reflejar en cada momento los propios del curso que se evalúa; por otro, las puntuaciones obtenidas deben estar en la misma métrica común para todas las aplicaciones. La violación de estos supuestos puede tener consecuencias importantes para la validez de los resultados (Gaviria y Ruiz de Miguel, 2007).

Dado que se trata de una evaluación longitudinal, las pruebas tienen que adaptarse al currículo de cada año. Los diseños de valor añadido se utilizan fundamentalmente para la realización de comparaciones a lo largo del tiempo, entre e intra centros. El garante de la «comparabilidad» dentro de los diseños longitudinales es el modelo de medida empleado que asume la métrica común de las escalas de medida. La posibilidad de la estimación en una escala común se basa en el supuesto de independencia local.

La conjugación de la complejidad curricular y el supuesto de independencia local se ven comprometidos cuando la evaluación abarca varias cohortes de distintas edades. La utilización de un modelo común de valor añadido en cohortes de niveles educativos distintos, por ejemplo Primaria y Secundaria, supone asumir previamente que se tiene la misma estructura dimensional de las pruebas en cada una de las cohortes.

Según Stout et al. (1996) la evaluación de la dimensionalidad de una prueba se da en dos tipos de problema: por un lado, la verificación o refutación de la unidimensionalidad y por otro, la consiguiente descripción de la estructura multidimensional de la prueba, si ésta fuese necesaria. La verificación de la unidimensionalidad es imprescindible en aquellos procedimientos psicométricos que presuponen que los datos se ajustan a un modelo latente unidimensional o, cuando menos, que el desajuste respecto de esta condición no invalida la utilización de un procedimiento específico.

Las soluciones más apropiadas según Martínez Arias et al. (2006) empiezan por llevar a cabo un análisis factorial lineal paramétrico sobre la matriz de correlaciones entre ítems, habitualmente sobre la matriz de correlaciones tetracóricas, analizándose la dimensionalidad en función de distintos criterios: valores propios superiores a la unidad, diagrama de sedimentación, cociente entre los dos primeros valores propios, etc. Pero también esta cuestión se puede abordar también mediante métodos no paramétricos, en los que no se parte de ningún supuesto sobre las funciones de respuesta. Los procedimientos HCA/CCPROX, DETECT y DIMTEST (disponibles en el programa DIMPACK) están basados en el concepto de unidimensionalidad esencial de Stout y en la estimación de las covarianzas condicionales de cada par de ítems.

Desde esta perspectiva, el objetivo de este trabajo es analizar, mediante dichos métodos no paramétricos, la estructura dimensional de las pruebas empleadas para evaluar el rendimiento académico en la asignatura de matemáticas en el contexto de la

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

evaluación realizada en la Comunidad de Madrid en los cursos académicos 2005-06 y 2006-07 en tres cohortes: 5º y 6º de Educación Primaria, 1º- 2º y 3º- 4º de ESO. Los datos utilizados para la realización de este trabajo proceden del Proyecto de I+D con referencia SEC2003-09742, ya finalizado, titulado: *El valor añadido en educación y la función de producción educativa: un estudio longitudinal.* Se ha evaluado la posible unidimensionalidad y la simplicidad o complejidad de la estructura de las pruebas.

Como se puede ver en la Tabla I, para los alumnos que en 2005-06 entraron en 5º de Primaria (cohorte 1ª) se han recogido datos en octubre de 2005 (primera medición), y junio de 2006 (segunda medición). En el curso 2006-07, cuando estos alumnos estaban en 6º de Primaria, se recogió información en noviembre de 2006 (tercera medición) y junio de 2007 (cuarta medición). Con las cohortes 2ª (correspondiente al primer ciclo de Secundaria) y 3ª (correspondiente al segundo ciclo de Secundaria) se ha procedido de manera análoga.

TABLA I. Tamaño de la muestra por cohortes y aplicaciones

Cohorte	Curso	Octubre 2005	Junio 2006	Curso	Noviembre 2006	Junio 2007
1ª	5º Primaria	4211	4173	6º Primaria	3627	3757
2ª	1º ESO	5106	4882	2º ESO	3327	3403
3ª	3º ESO	4736	4272	4º ESO	2761	2772

Las variables dependientes utilizadas han sido las puntuaciones obtenidas en las pruebas de Matemáticas elaboradas *ad hoc* aplicadas a los alumnos de los centros de la muestra.

Análisis mediante métodos no paramétricos basados en la TRI

Introducción

Como al inicio se ha mencionado, además de los enfoques basados en las técnicas factoriales, la evaluación de la dimensionalidad de las pruebas es posible abordarla también con procedimientos no paramétricos basados en la Teoría de Respuesta al Ítem. En concreto aquí vamos a presentar la aplicación a nuestras pruebas de los procedimientos DETECT y DIMTEST que forman parte del paquete DIMPACK elaborado por

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

The Roussos-Stout Software Development Group y distribuido por Assesment Systems Corp (2006).

Estos procedimientos se basan en el concepto de «unidimensionalidad esencial» (Stout, 1987, 1990) y en la covarianza condicional de pares de ítems (Stout et al., 1996). Este principio de dimensionalidad esencial se basa a su vez en la distinción establecida por McDonald (1981) entre los principios «fuerte» y «débil» de independencia local (Strong Local Independence, SLI; Weak Local Independence, WLI).

Stout define la dimensionalidad esencial de una prueba como el número mínimo de dimensiones del espacio de rasgos latentes tal que produzca un modelo que sea localmente independiente y monótono. Conceptualmente, la idea es que una prueba es «esencialmente unidimensional» si la puntuación de la misma está en función de «un» rasgo latente dominante y también de uno o varios rasgos menores, secundarios o de ruido *(nuisance latent traits).* Una descripción más formalizada se encuentra en Zhang y Stout (1999), y en Jang y Roussos (2007).

Procedimiento DETECT

DETECT es un procedimiento cuya finalidad es estudiar la posible multidimensionalidad de una prueba. Para ello obtiene el índice DETECT que expresa la «cantidad» de multidimensionalidad de la misma. Esto, en términos geométricos, hace referencia a la dispersión de los vectores de los ítems. En cualquier caso, no debe ser confundido con el número de dimensiones de una prueba. Este procedimiento está basado en que, dada una prueba multidimensional, los ítems que miden la misma dimensión tendrán covarianzas condicionales positivas, mientras que para los que midan otras dimensiones distintas, sus covarianzas serán negativas de forma que el citado índice puede ser considerado como una media ponderada (Stout, W., Nandakumar, R. y Habing, R., 1996, Roussos y Ozbek, 2006 y Stout et al., 2006). Para una descripción rigurosa de la fundamentación matemática de este índice puede recurrirse al trabajo de Zhang (2007).

El procedimiento DETECT puede emplearse tanto desde un planteamiento exploratorio como confirmatorio. En el primer modo, DETECT busca cuál es la partición que maximiza el índice y asigna ítems a los conglomerados. En modo confirmatorio una submuestra se emplea para la parte exploratoria y otra para validar la partición, es decir, para validar el modelo propuesto por el usuario (Jang y Roussos, 2007). En general, si se dispone de muestras de tamaño suficiente, se recomienda usar el procedimiento de validación cruzada (Monahan et al., 2007).

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

Siguiendo estas pautas, hemos aplicado el procedimiento DETECT a las 24 pruebas que estamos estudiando después de haber dividido la muestra total en dos submuestras de similar número de efectivos con objeto de que una opere como submuestra de exploración y la otra de validación. En la Tabla II, que presentamos a continuación, aparecen los resultados de este proceso.

TABLA II. Aplicación de DETECT con validación cruzada a las 24 pruebas

Cohorte	Octubre 2005		Junio 2006		Octubre 2006		Junio 2007		Medias
	A	B	A	B	A	B	A	B	
	DETECT:	DETECT:	DETECT:	DETECT:	DETECT:	DETECT:	DETECT:	DETECT:	
	0.0892	0.0966	0.1261	0.1332	0.1997	0.2398	0.1113	0.1575	0,1442
	IDN :	IDN :	IDN :	IDN :	IDN :	IDN :	IDN :	IDN :	
5º-6º	0.5513	0.5385	0.5692	0.5756	0.5775	0.5978	0.5377	0.5718	0,5649
	Ratio:	Ratio:	Ratio:	Ratio:	Ratio:	Ratio:	Ratio:	Ratio:	
	0.2234	0.2215	0.2270	0.2663	0.3058	0.3577	0.1918	0.2306	0,2530
	Max: 4	Max: 4	Max: 4	Max: 3	Max: 4	Max: 4	Max: 4	Max: 4	3,8750
	DETECT:	DETECT:	DETECT:	DETECT:	DETECT:	DETECT:	DETECT:	DETECT:	
	0.1137	0.2034	0.1187	0.1762	0.1686	0.0926	0.1567	0.0993	0,1412
	IDN :	IDN :	IDN :	IDN :	IDN :	IDN :	IDN :	IDN :	
1º-2º	0.5601	0.6248	0.5548	0.5968	0.5965	0.5439	0.5526	0.5577	0,5734
	Ratio:	Ratio:	Ratio:	Ratio:	Ratio:	Ratio:	Ratio:	Ratio:	
	0.2084	0.3468	0.1895	0.3007	0.2610	0.1367	0.2312	0.1472	0,2277
	Max: 3	Max: 5	Max: 4	Max: 4	Max: 5	Max: 4	Max: 5	Max: 4	4,2500
	DETECT:	DETECT:	DETECT:	DETECT:	DETECT:	DETECT:	DETECT:	DETECT:	
	0.2966	0.1205	0.2292	0.1568	0.2337	0.2426	0.1683	0.1662	0,2017
	IDN :	IDN :	IDN :	IDN :	IDN :	IDN :	IDN :	IDN :	
3º-4º	0.6628	0.5692	0.6323	0.5833	0.5868	0.5946	0.5505	0.5420	0,5902
	Ratio:	Ratio:	Ratio:	Ratio:	Ratio:	Ratio:	Ratio:	Ratio:	
	0.4968	0.2009	0.4947	0.2529	0.2979	0.2739	0.1947	0.2041	0,3020
	Max: 4	Max: 4	Max: 5	Max: 4	Max: 4	Max: 4	Max: 4	Max: 5	4,2500

El contenido de cada casilla es como sigue:

- DETECT: es el índice del mismo nombre y ha de interpretarse como un indicador de la cantidad de multidimensionalidad. Inmediatamente vamos a exponer los criterios de interpretación de este índice.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

- IDN: es un índice que informa del grado de simplicidad o complejidad de la estructura, de forma que valores cercanos a 1 denotan una estructura simple.
- Ratio: es la proporción entre el valor del índice DETECT obtenido y su valor máximo posible. Al igual que en el caso anterior, cuanto más cercano a 1, más simple se puede considerar la estructura.
- Max: indica el número de particiones de los ítems que hacen máximo el valor del índice DETECT.

Veamos con más detalle cada uno de estos parámetros. Como en su momento se apuntó, el procedimiento comienza buscando el número de particiones que hacen máximo el índice DETECT. El valor apuntado, valorado conjuntamente con los demás, es un buen indicador del tipo de estructura con que nos encontramos. Así vemos que en nuestro caso, en la gran mayoría de las pruebas (17) el número de particiones que maximizan el índice es de cuatro, en cinco casos son cinco; y sólo en dos el máximo está en tres. El índice DETECT es una media ponderada de todas las covarianzas condicionales de forma que el peso será +1 cuando los dos ítems provengan de un mismo conglomerado, y -1 si la pareja procede de distintos conglomerados (Jang y Roussos, 2007). En relación a los criterios de interpretación del mismo, en la literatura nos encontramos con varios, aunque básicamente se pueden resumir en los dos que presentamos en la Tabla III (Monahan et al., 2007).

TABLA III. Tamaño de la muestra por cohortes y aplicaciones

Unidimensionalidad esencial	0,0 - 0,1		
Multidimensionalidad débil	0,1 - 0,5	Unidimensionalidad esencial	0,0 - 0,2
Multidimensionalidad moderada	0,5 - 1,0	Multidimensionalidad de débil a moderada	0,2 - 0,4
Multidimensionalidad fuerte	1,0 - 1,5	Multidimensionalidad de moderada a grande	0,4 - 1,0
Multidimensionalidad muy fuerte	Sup. 1,5	Multidimensionalidad muy fuerte	Sup. 1,0

En nuestro caso, el optar por uno u otro no es baladí porque, como puede verse, la mayoría de las pruebas obtienen valores que se encuentran en los límites de ambos criterios (17 pruebas tienen un índice con un valor comprendido entre 0,11 y 0,18; y el promedio de las 24 es 0,1624) lo que supone que, si se adopta el primer criterio, la mayoría de las pruebas podrían considerarse como esencialmente unidimensionales; mientras que si optamos por el segundo, hablaríamos más bien de una débil multidimensionalidad.

Con respecto al segundo índice, el valor del IDN se basa en la proporción de signos positivos y negativos de las covarianzas condicionales con respecto al máximo

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

posible que se da cuando todos los signos «intra-conglomerados» son positivos y todos los «entre-conglomerados» son negativos. Este índice ha de interpretarse, por tanto, como un indicador de hasta qué punto la estructura puede considerarse como simple o compleja. Se entiende como estructura simple perfecta aquella en que los ítems se agrupan en k dimensiones que son subconjuntos disjuntos. Obviamente no existen estructuras simples perfectas (que darían un valor de uno en este índice), sino que esta característica se puede considerar como un continuo que va desde cero (máxima complejidad) hasta uno (máxima simplicidad).

Y con respecto al tercer parámetro, ya hemos anticipado que es la ratio entre el valor del índice DETECT efectivamente obtenido y el máximo que puede alcanzar. Su interpretación es similar al anterior.

A modo de resumen final, del examen de los promedios por cohortes podemos concluir que –en conjunto– nos encontramos con unas pruebas de multidimensionalidad débil o incluso en algunos casos esencialmente unidimensionales. Si aplicamos el criterio más riguroso nos encontramos con 4 de ellas unidimensionales (índice menor de 0,1). Junto a éstas hay 14 cuyo valor está comprendido entre 0,10 y 0,20 lo que supone, dependiendo del criterio que se emplee, que se calificarían bien como unidimensionales, bien de débil multidimensionalidad.

Lo que sí hay que señalar es que el comportamiento de este índice no es igual en las tres cohortes pues la de 3º- 4º tiene una media claramente superior. En lo que no hay tanta variación (aunque la tercera cohorte también arroja el valor más alto) es en lo relativo al índice de complejidad cuyo promedio global es 0,576 lo que nos indica que –al margen de que las pruebas sean o no esencialmente unidimensionales– la estructura es de una cierta complejidad. Este resultado es también claramente coincidente con la estructura factorial que hemos encontrado en la que, como vimos, para todas las pruebas, prácticamente la gran mayoría de los ítems carga en el primer factor; mientras que, con respecto al segundo, son pocos los ítems que saturan en el mismo sin hacerlo simultáneamente en el primero.

Esto parece indicar que nos encontramos ante una estructura dimensional compuesta muy probablemente por una dimensión mayor, principal, y una o varias secundarias o menores con incluso algunas compuestas por sólo un ítem. Es una estructura típica que Stout caracterizó al formular la unidimensionalidad esencial señalando además –como ocurre en nuestro caso– que «una estructura simple no implica que las dimensiones no estén correlacionadas» (Stout, 1996, p. 332). Como ahora veremos, el análisis de conglomerados jerárquico apunta también en este sentido.

Procedimiento DIMTEST

El procedimiento DIMTEST (Stout et al., 1996; Nandakumar y Stout, 1993; Stout, 1987) se basa en comparar la covarianza condicional de dos subconjuntos de ítems de una prueba (el subconjunto de evaluación, AT; y el de partición, PT). Para ello contrasta la hipótesis nula de que AT es de similar dimensionalidad que PT, frente a la hipótesis alternativa de que AT es dimensionalmente homogéneo y distinto que PT.

El principio de esta comparación radica en que las covarianzas condicionales de DIMTEST son las covarianzas entre los ítems de AT condicionadas al rasgo latente θ siendo medido por PT. La idea es que si los ítems de AT y los de PT miden la misma dimensión, entonces las covarianzas condicionales entre los ítems de AT serán igual a cero, mientras que si AT mide una dimensión distinta que PT, entonces las covarianzas condicionales serán positivas (Jang y Roussos, 2007). El contraste se lleva a cabo mediante el estadístico T, y los detalles relativos al mismo tanto en la versión inicial que requería dos subconjuntos AT como en la actual que sólo emplea uno, se pueden encontrar en Stout, Froelich y Gao (2001, p. 5-14) y en Stout, Nandakumar y Habing (1996, p. 44-47).

Este principio así formulado hace que la elección de los ítems que conforman AT sea crucial y que determine en gran medida el resultado de la prueba estadística. Esto ya fue observado por Wang y Hocevar (1994) en un trabajo en el que precisamente estudiaban la dimensionalidad de pruebas de matemáticas cuando afirmaron que el grado de dimensionalidad esencial depende de la elección de ítems de AT. A este respecto, los criterios de selección de los ítems que conforman AT son formulados de la siguiente manera (Stout et al., 1996; Stout, Froelich y Gao, 2001):

- Los ítems de AT han de ser relativamente homogéneos.
- Los ítems de AT han de ser lo más distintos con respecto a los de PT.
- El número de ítems de PT ha de ser lo suficientemente grande (al menos 15), y el de AT no más de un tercio de los de PT con un mínimo de tres.

Es decir, dado el contraste que DIMTEST realiza, se trata de seleccionar como miembros pertenecientes al subconjunto AT aquellos ítems que conformen un aglomerado lo más unidimensional posible y en una dimensión distinta del resto de los ítems de la prueba (que conforman PT). De esta manera si, a priori, la selección es exigente y AT es lo más unidimensional posible y lo más distinto posible de PT, podemos estar razonablemente seguros de que si el valor de T tiene una probabilidad mayor

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

que el nivel alfa, si se acepta la hipótesis nula, entonces la prueba en su conjunto se puede considerar como esencialmente unidimensional dado que a pesar de haber maximizado a priori la diferencia entre ambos subconjuntos, la misma no ha resultado finalmente significativa.

Y los métodos propuestos para encontrar ítems que satisfagan estos criterios son el análisis factorial exploratorio, el análisis de conglomerados jerárquico y el juicio de expertos basado en el contenido, el dominio cognitivo y la estructura de la prueba. Y en esta fase del trabajo, nosotros hemos aplicado estos tres criterios. Veamos los resultados de cada uno de ellos.

Básicamente DIMTEST ofrece dos procedimientos para seleccionar los ítems de AT: o son definidos por el usuario en función de algún criterio de los anteriormente señalados, o bien se emplea un algoritmo incorporado que realiza esta tarea por nosotros. En primer lugar, probamos a emplear como criterio los conglomerados generados por DETECT o por HCC. En el primer caso, como ya comentamos, existe la posibilidad de especificar como parámetro del procedimiento el número de dimensiones, de conglomerados a extraer. Pues bien, si este valor lo fijamos en dos, obtenemos una partición de los ítems en dos subconjuntos disjuntos que podemos emplear (en función de su tamaño relativo) como AT y PT. De manera similar, si para cada prueba realizamos un análisis de conglomerados jerárquico mediante el procedimiento HCC y seleccionamos el penúltimo paso (la solución con dos conglomerados) podemos operar de la misma forma.

Pues bien, en ambos casos y para la mayoría de las 24 pruebas, los resultados de DIMTEST son los que cabía esperar: para aquellas pruebas en que DETECT había arrojado un valor del índice que denotaba unidimensionalidad esencial, el contraste de T mediante DIMTEST resultaba no significativo, mientras que en las que el índice era alto, la prueba resultaba estadísticamente significativa. Por último, para los valores intermedios del índice DETECT (de dimensionalidad débil), los resultados variaban.

Desde este punto de vista, podemos considerar que, en nuestro caso, el empleo de DIMTEST después de DETECT o HCC resultaba redundante y no nos aportaba información adicional relevante.

A continuación, optamos por que fuera el propio programa el que realizase la selección de los ítems de AT. El criterio de selección que el programa incorpora emplea los resultados de un análisis exploratorio de componentes principales, de forma que, de la solución no rotada, selecciona aquellos ítems cuyas saturaciones son altas en el segundo factor y bajas en el primero. Esto en principio es un procedimiento muy coherente con la estrategia analítica de DIMTEST. Véase Tabla IV.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

TABLA IV. Resultados DIMTEST, método automático AFE

P (AFE)	Octubre 2005		Junio 2006		Octubre 2006		Junio 2007	
	A	B	A	B	A	B	A	B
5°-6°	0.0032	0.0125	0.0001	0.0000	0.0000	0.0000	0.1387	0.0190
1°-2°	0.0046	0.0010	0.0000	0.0000	0.4708	0.0000	0.4434	0.1412
3°-4°	0.0000	0.0015	0.0485	0.0000	0.0000	0.0036	0.0075	0.0002

Como vemos, los resultados aquí podrían, en principio, considerarse muy divergentes con respecto a los obtenidos mediante DETECT. Aquí sólo seis de las 24 pruebas resultan esencialmente unidimensionales, y eso empleando como nivel alfa el 0,05. Si optamos por el 0,01 se nos quedan en cuatro. Por el contrario, recordemos cómo eran 18 las pruebas que tenían un índice DETECT inferior a 0,2. Así vistos los resultados, la divergencia entre ambos procedimientos parece grande, pero la cuestión toma otro cariz si en el caso de DETECT optamos por un criterio más exigente, pues entonces sólo son cuatro las pruebas con índice inferior a 0,10. Desde esta perspectiva parecería que los enfoques convergen, aunque si examinamos las tablas con detalle veremos que es sólo a nivel numérico global pues son pocas las pruebas cuyos resultados coinciden bajo ambos criterios (más adelante resumiremos esto en una tabla global).

En nuestra opinión, esta divergencia entre ambos enfoques puede explicarse si nos detenemos a examinar el modo en que este criterio «automático» de DIMTEST procede. Como hemos dicho, se seleccionan los ítems del segundo factor no rotado de ACP. Y lo que ocurre es que, como vimos antes, la gran mayoría de estas pruebas tienen una estructura compleja y esto hace que el criterio adoptado por DIMTEST no funcione bien pues no hay casi ítems que claramente carguen en el segundo factor sin hacerlo simultáneamente en el primero. Este hecho es señalado por Finch (2006) al indicar que habitualmente las soluciones iniciales no reflejan una estructura simple. Con objeto de afrontar este problema, hemos optado por un criterio que emplee las soluciones factoriales, pero no del ACP sin rotar, sino los resultados del análisis factorial de información completa con solución rotada según el criterio PROMAX que anteriormente hemos expuesto.

Pensamos que este enfoque permite satisfacer mejor los criterios que DIMTEST recomienda a la hora de seleccionar ítems para AT pues la solución rotada permite seleccionar ítems que sólo carguen en el segundo factor sin hacerlo el primero. Es cierto que al tratarse de una solución oblicua, ambos factores están relacionados, pero es que la estructura de las pruebas parece indicar la existencia de una dimensión principal que estaría representada por el primer factor y una dimensión secundaria de mayor o menor importancia relativa en las diferentes pruebas que sería el segundo factor. (Hemos probado también soluciones con rotación ortogonal como VARIMAX

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

pero ocurre prácticamente lo mismo que con las soluciones iniciales no rotadas). En concreto hemos procedido de la siguiente manera.

En primer lugar procedimos para cada prueba a la selección de los ítems de AT empleando las cargas factoriales de la solución PROMAX obtenida mediante TESTFACT. En aplicación de los criterios de DIMTEST, se seleccionaron para cada prueba aquellos ítems cuyas cargas factoriales fuesen iguales o superiores a 0,35 en el segundo factor y menores que 0,15 en el primero (Finch, 2006).

Las Tablas V y VI muestran respectivamente los valores de la probabilidad, y el número, contenido y dominio cognitivo de los ítems seleccionados como miembros de AT en aplicación del criterio anterior.

TABLA V. Resultados DIMTEST, AT manual mediante saturaciones PROMAX

P (Promax)	Octubre 2005		Junio 2006		Octubre 2006		Junio 2007	
	A	B	A	B	A	B	A	B
5º - 6º	0.4006	0.0789	0.0008	0.0001	0.1236	0.0000	0.3901	0.2039
1º - 2º	0.2849	0.1074	0.0180	0.0212	0.0002	0.0145	0.0905	0.2493
3º - 4º	0.0002	0.0000	0.0782	0.0234	0.0191	0.0019	0.1562	0.0007

TABLA VI. Descripción ítems solución anterior

Ítems		Octubre 2005		Junio 2006		Octubre 2006		Junio 2007	
		A	B	A	B	A	B	A	B
5º-6º	Con.	Nu:4	Nu:4	Nu:0	Nu:0	Nu:6	Nu:0	Nu:3	Nu:1
		Ge:0	Ge:0	Ge:4	Ge:5	Ge:1	Ge:7	Ge:1	Ge:2
		Es:0	Es:0	Es:0	Es:0	Es:1	Es:2	Es:2	Es:2
		Me:0	Me:0	Me:0	Me:1	Me:3	Me:0	Me:1	Me:2
	Dom.	Con:1	Con:2	Con:4	Con:5	Con:5	Con:4	Con:5	Con:3
		Apr:3	Apr:2	Apr:0	Apr:1	Apr:6	Apr:5	Apr:2	Apr:4
	N	4	4	4	6	11	9	7	7
1º-2º	Con.	Nu:3	Nu:1	Nu:4	Nu:8	Nu:7	Nu:1	Nu:5	Nu:2
		Ge:1	Ge:2	Ge:3	Ge:3	Ge:0	Ge:4	Ge:1	Ge:1
		Me:0	Me:0	Me:0	Me:0	Me:0	Me:1	Me:0	Me:0
		Ta:1	Ta:0	Ta:0	Ta:1	Ta:0	Ta:0	Ta:0	Ta:1
	Dom.	Con:4	Con:1	Con:2	Con:5	Con:6	Con:2	Con:1	Con:2
		Apr:1	Apr:2	Apr:5	Apr:7	Apr:1	Apr:4	Apr:5	Apr:2
	N	5	3	7	12	7	6	6	4
3º-4º	Con.	Nu:2	Nu:4	Nu:0	Nu:2	Nu:7	Nu:2	Nu:1	Nu:3
		Ge:0	Ge:0	Ge:0	Ge:0	Ge:0	Ge:1	Ge:4	Ge:0
		Fu:1	Fu:0	Fu:3	Fu:0	Fu:0	Fu:2	Fu:1	Fu:3
		Es:0	Es:1	Es:0	Es:1	Es:0	Es:2	Es:0	Es:0
	Dom.	Con:1	Con:2	Con:0	Con:1	Con:2	Con:0	Con:0	Con:3
		Apr:2	Apr:3	Apr:3	Apr:2	Apr:5	Apr:7	Apr:6	Apr:3
	N	3	5	3	3	7	7	6	6

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

La solución aportada por DIMTEST para estos AT es la que finalmente adoptamos y del examen de la misma vemos que nos encontramos con once pruebas que podemos considerar como básicamente unidimensionales. Tal y como antes comentamos, al final de este trabajo compararemos todos los criterios y procedimientos que hemos empleado con objeto de triangular y proponer una solución final global, por lo que ahora nos vamos a limitar a señalar cómo las dos primeras cohortes tienen una proporción mayor de pruebas unidimensionales, mientras que la cohorte de 3º y 4º de la ESO se puede considerar como no unidimensional lo que de nuevo viene a confirmar lo comentado anteriormente.

Pero ahora nos interesa centrarnos en los ítems seleccionados como miembros de AT. Por término medio, la aplicación de los criterios señalados supone que el tamaño medio de este subconjunto es de 5,8 ítems (5,45 para las pruebas unidimensionales y 6,3 en el caso de las otras). En consecuencia, la proporción AT-PT es de aproximadamente uno a seis o siete dado que el número de ítems de las pruebas oscila entre 36 y 40.

Estos valores pensamos que son un buen indicador del tipo de estructura que nos encontramos, pero además de esto nos interesa también fijarnos en el contenido sustantivo de los mismos. Según la matriz de especificaciones elaborada por los diseñadores de las pruebas (inspectores), los contenidos y dominios cognitivos son los que aparecen en la Tabla VI. Los contenidos son: números, geometría, estadística, medida, tablas y funciones; y los dominios cognitivos son conocimientos (Con) y aplicación o resolución de problemas (Apr). En todas las pruebas los ítems mayoritarios pertenecen al contenido «numérico-aritmético» y al dominio cognitivo de conocimientos.

Y del examen de la Tabla VI podemos ver que en gran parte de los casos nos encontramos que este subconjunto AT está formado mayoritariamente bien por ítems de contenidos distintos al numérico (geometría, estadística, medida, etc.) y/o por aquellos ítems que suponen aplicación o resolución de problemas. Esta tendencia no es unánime ni homogénea pero sí mayoritaria tanto para las pruebas unidimensionales como para las restantes. Vamos a ver con algo de detalle un caso paradigmático como es el de la forma B de la prueba aplicada a 6º de Primaria en octubre de 2006, pero no hay que olvidar que junto a un caso tan claro como éste hay notables excepciones que al final enumeraremos.

El valor de T asociado a esta prueba es de 6,64 lo que supone una probabilidad menor que 0,0001. Es decir, nos encontramos con una prueba claramente no unidimensional desde la perspectiva de DIMTEST. Los nueve ítems que conforman el subconjunto AT (1, 3, 5, 6, 7, 8, 9, 10 y 11) son dos de estadística y nueve de geometría, mientras que desde el punto de vista del dominio se reparte a medias entre conocimientos

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

y resolución de problemas. Los 30 ítems restantes (que componen PT) son mayoritariamente de contenido numérico (16) o relacionados con la medida y las magnitudes (diez) habiendo también tres de geometría y una de estadística.

Para mayor evidencia en este sentido, nos encontramos además que uno de los conglomerados propuestos por el procedimiento DETECT está compuesto básicamente por estos ítems y lo mismo ocurre con la solución de cinco conglomerados de HCC en que éstos conforman el quinto grupo. (A este respecto, pensamos que de nuevo es síntoma de complejidad estructural el hecho de que sea en el quinto grupo cuando surja esta estructura y no en la de dos conglomerados lo que nos ratifica en lo que antes apuntamos: dada esta complejidad las soluciones de dos conglomerados proporcionadas por DETECT o HCC no son satisfactorias).

En conclusión, parece claro en este caso que nos encontramos con una prueba no unidimensional compuesta por una primera dimensión principal en la que saturan la mayoría de los ítems y una secundaria centrada fundamentalmente en ítems de geometría.

Como antes apuntábamos, esta caracterización de los dos factores puede encontrarse en muchas de las pruebas (sean o no esencialmente unidimensionales), pero hay también excepciones. Y entre ellas destacamos la forma B de 5º de primaria aplicada en octubre de 2005, la forma A de 6º de primaria de octubre de 2006 (la que debería ser equivalente a la que acabamos de describir y que al menos dimensionalmente no lo es), la forma A de 6º de primaria aplicada en junio de 2007, la forma A de 1º de la ESO de octubre de 2005 y, por último, la forma A de 2º de la ESO de octubre de 2006. Queremos resaltar el caso de ésta última porque, además de no ajustarse al patrón descrito, es además no unidimensional. Se trata de cinco claras excepciones sobre un total de 24, por lo que la existencia de un único patrón debe formularse con las debidas precauciones.

Stout, Nandakumar y Habing (1996); Perkhounkova y Dunbar (1999); y Walker et al. (2006), al analizar pruebas de matemáticas llevan a cabo planteamientos similares combinando el análisis de contenido de ítems y pruebas con los resultados de DIMTEST. Hattie et al. (1996) concluyen su trabajo afirmando que DIMTEST está basado en el principio débil de independencia local y está diseñado, no tanto para identificar si un conjunto de ítems es o no unidimensional, sino si hay una dimensión suficientemente dominante. Si además combinamos las perspectivas que nos proporcionan DETECT y DIMTEST pensamos que es posible lograr evidencias relativas a la posible unidimensionalidad esencial de las pruebas en el sentido que aquí hemos concluido.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

Comparación de resultados y conclusiones finales

A lo largo de este trabajo hemos abordado la dimensionalidad de las 24 pruebas desde diferentes perspectivas y con diferentes procedimientos y herramientas analíticas. En los distintos apartados hemos ido apuntando las conclusiones parciales y hemos ido igualmente señalando las concordancias entre criterios así como también las divergencias observadas. Para finalizar, hora es que recopilemos sintetizando resultados y apuntando las conclusiones finales.

Con objeto de evaluar la dimensionalidad de las pruebas, seis son los principales criterios que hemos empleado:

- Que en la solución factorial, sólo tenga un autovalor superior a uno el primer factor.
- Que el cociente entre los autovalores de los dos primeros factores sea superior a cinco.
- La significatividad de la T según el procedimiento DIMTEST habiendo seleccionado los ítems de AT según las cargas factoriales de la solución PROMAX.
- La significatividad de la T según el procedimiento DIMTEST habiendo seleccionado los ítems de AT el propio programa mediante AFE.
- Que el valor del índice DETECT sea inferior a 0,10.
- Que el valor del índice DETECT sea inferior a 0,15.

Pues bien, en la Tabla VII, siguiendo la presentación habitual de cohortes, aplicaciones y formas de las pruebas, se consigna en cada casilla los criterios que cada prueba satisface.

TABLA VII. Síntesis de criterios

Examen	Octubre 2005		Junio 2006		Noviembre 2006		Junio 2007	
Conjunto	A	B	A	B	A	B	A	B
5° - 6°	3,5,6	3,4,5,6	2,6	2,6	2,3	2	1,2,3,4,6	1,2,3,4
1° - 2°	1,2,3,6	3	1,2,6	1,2	4	5,6	2,3,4	3,4,5,6
3° - 4°		1,6	3,4		1	1	1,3	1

A la vista de esto hemos clasificado las 24 pruebas en tres categorías atendiendo a su unidimensionalidad esencial y nos encontramos con:

- Diez pruebas de las que parece haber evidencia suficiente de unidimensionalidad esencial (casillas sombreadas en gris oscuro).

- Seis pruebas de multidimensionalidad débil o en las que la evidencia de unidimensionalidad esencial no es tan concluyente (casillas sombreadas en gris claro).
- Ocho pruebas de multidimensionalidad baja o moderada (casillas sin sombreado).

Los criterios para asignar a cada una de estas tres categorías son:

- Se consideran pertenecientes al primer grupo (esencialmente unidimensionales) si:
  - el valor del índice DETECT tiene un valor inferior a 0,10.
  - si DETECT es inferior a 0,15 y además satisfacen otros dos criterios más cualesquiera.
  - si el valor de T no ha resultado significativo en los 2 procedimientos DIMTEST.
- Se consideran de multidimensionalidad débil aquellos que satisfacen uno de los criterios tres o seis.
- Y el resto son considerados de multidimensionalidad baja o moderada.

Como vemos, la prelación de los criterios es clara. Si el índice DETECT es inferior a 0,10 la prueba se considera esencialmente unidimensional. La misma categoría tienen aquellas pruebas cuyo índice es menor que 0,15 y que además hayan satisfecho otros dos criterios. Y lo mismo ocurre si en los dos procedimientos DIMTEST, la T de Stout ha resultado no significativa. En definitiva, el criterio básico es el índice DETECT en su versión más estricta y en los demás casos se considera esencialmente unidimensional si se han obtenido evidencias adicionales satisfaciendo dos (si son DIMTEST) ó más de dos criterios. En este primer grupo se encuentran diez pruebas que como vemos pertenecen cuatro a la primera cohorte, cinco a la segunda y sólo una a la tercera. En el grupo intermedio se encuentran las seis pruebas que sólo han satisfecho dos criterios siendo éstos el dos o el tres. Y el tercer grupo es aquel en el que hay pocas o ninguna evidencia de unidimensionalidad. Se trata de ocho pruebas de las cuáles cinco pertenecen a la tercera cohorte. En general, son pruebas de baja o moderada multidimensionalidad pues no olvidemos que el índice DETECT es muy bajo en general. De las ocho pruebas no todas siguen un patrón idéntico:

- Podemos considerar como las de una multidimensionalidad relativa más clara las pruebas correspondientes a las aplicaciones de octubre 2005 (forma A), las dos formas de octubre de 2006 y la forma B de junio de 2007, todas ellas de la tercera cohorte.

- La aplicación de la forma A de octubre de 2006 de 2º de la ESO es peculiar pues resulta unidimensional en el procedimiento DIMTEST con AFE (criterio cuatro) con una probabilidad muy elevada, y en cambio sólo resulta unidimensional en dicho criterio.
- Todas las pruebas marcan en algún criterio, excepto la forma A de octubre 2005 y la B de junio de 2006 de la tercera cohorte.

Si examinamos la Tabla VIII por filas (cohortes), vemos que la primera (5º y 6º de primaria) resulta ser la más unidimensional, seguida de la de 1º y 2º de la ESO; y es, como ya veníamos comentando, la de 3º y 4º de la ESO la de multidimensionalidad más acusada.

TABLA VIII. Categorías de dimensionalidad

	Esencialmente unidimensional	Multidimensionalidad. débil	Baja o moderada multidimensionalidad
5º-6º	4	3	1
1º-2º	5	1	-
3º-4º	1	2	5

De los seis criterios estudiados, el menos fiable es el primero pues son nueve las pruebas que marcan en él y de las mismas, cuatro resultan de la tercera categoría, dos de la segunda y tres de la primera. En definitiva, no parece discriminar adecuadamente.

Por último, la Tabla IX muestra la simplicidad o complejidad relativa de las pruebas en función del índice IDN proporcionado por DETECT. Se han considerado de estructura simple (y sombreado en gris) aquellas pruebas con un índice superior a 0,58. Dentro de las mismas se ha consignado si, de acuerdo con la síntesis anterior, resultaban multidimensionales. Vemos como de las nueve que resultan de estructura más relativamente simple, siete pertenecen a la tercera categoría.

TABLA IX. Simplicidad-complejidad relativa

Estructura simple	Octubre 2005		Junio 2006		Noviembre 2006		Junio 2007	
vs. Compleja	A	B	A	B	A	B	A	B
5º - 6º						Multidim.		
1º - 2º				Multidim.	Multidim.			
3º - 4º	Multidim.			Multidim.	Multidim.	Multidim.		

Conclusiones finales

Si a todo esto unimos a la caracterización de los dos factores que hemos ido desarrollando, para acabar podemos concluir –con las matizaciones que hemos ido apuntando– lo siguiente:

- La gran mayoría de las pruebas de las dos primeras cohortes se pueden considerar esencialmente unidimensionales o de débil multidimensionalidad.
- Por el contrario, las pruebas de la tercera cohorte son mayoritariamente de moderada multidimensionalidad (en términos relativos).
- La mayoría de las pruebas tienen una estructura de una cierta complejidad y las que resultan ser de estructura más simple tienden a mayor multidimensionalidad.
- Esto nos reafirma en la idea de que, mayoritariamente, nos encontramos con una dimensión latente principal de la que participan la gran mayoría de los ítems y una secundaria.
- En las pruebas de estructura simple, podría hablarse probablemente de dos dimensiones más marcadas. En la primera seguirían cargando muchos ítems, y la segunda, más claramente definida, se aglutinaría alrededor de contenidos como la geometría, la medida o la estadística o por dominios cognitivos de aplicación o resolución de problemas.
- Para las tres cohortes y desde la perspectiva de la dimensionalidad, las formas A y B no resultan equivalentes.
- Desde una perspectiva longitudinal, en las dos primeras cohortes la estructura dimensional de las pruebas se mantiene básicamente a lo largo de las cuatro aplicaciones. No ocurre lo mismo con la tercera.
- La triangulación realizada mediante los procedimientos factoriales «clásicos», el análisis factorial de información completa y los procedimientos no paramétricos basados en TRI ha resultado una estrategia metodológica válida que nos ha permitido acumular evidencia empírica que consideramos suficiente para apoyar las conclusiones anteriores.

Todos estos resultados confirman que el aumento de la complejidad de los contenidos implica un aumento en la complejidad de la estructura dimensional de las pruebas. Los modelos de medida utilizados son ciertamente bastante robustos a las violaciones leves de los supuestos relacionados con la independencia local.

En el futuro, siempre que la evaluación de valor añadido se refiera a un abanico de cursos tan amplio como el que aquí se ha estudiado, se recomienda tener muy en

cuenta la posibilidad de contar con estructuras de dimensionalidad de complejidad creciente en paralelo con el desarrollo del currículo.

Referencias bibliográficas

Abad, F. J., Ponsoda, V. y Revuelta, J. (2006). *Modelos politómicos de respuesta al ítem.* Madrid: La Muralla.

Finch, H. (2006). Comparison of the Performance of Varimax and Promax Rotations: Factor Structure Recovery for Dichotomous Ítems. *Journal of Educational Measurement,* 43, 39–52.

Gaviria, J. L. y Ruiz de Miguel, C. (2007). Importancia de algunos supuestos psicométricos en la evaluación de los sistemas educativos. Calibración y equiparación en las pruebas de Estándares Nacionales de México. *Revista de Educación,* 343, 223-248.

Hattie, J., Krakowski, K., Rogers H.J. & Swaminathan, H. (1996). An Assessment of Stout's Index of Essential Unidimensionality. *Applied Psychological Measurement.* 20, 1-14.

Jang, E.E. & Roussos, L. (2007). An Investigation into the Dimensionality of TOEFL Using Conditional Covariance-Based Nonparametric Approach. *Journal of Educational Measurement,* 44, 1, 1–21.

Martínez, M. R., Hernández, M. J. y Hernández, M. V. (2006). *Psicometría.* Madrid: Alianza.

Monahan, P.O., Stump, T.E., Finch, H. & Hambleton, R.K. (2007). Bias of Exploratory and Cross-Validated DETECT Index Under Unidimensionality. *Applied Psychological Measurement,* 31, 483-503.

Nandakumar, R. & Stout, W. (1993). Refinements of Stout's procedure for assessing unidimensionality. *Journal of Educational Statistics.* 18, 41-68.

Roussos, L.A. & Ozbek, O.Y. (2006). Formulation of the DETECT Population Parameter and Evaluation of DETECT Estimator Bias. *Journal of Educational Measurement,* 43, 3, 215-243.

Roussos, L.A. & Stout, W. (2007). *Dimpack. Version 1.0.* The Roussos-Stout Software Development Group. St. Paul, MN: Assesment Systems Corp.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika,* 52, 589-617.

— (1990). A new item response theory modelling approach with applications to unidimensional assessment and ability estimation. *Psychometrika,* 55, 293-326.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

Stout, W., Habing, B., Douglas, J., Kim, H., Roussos, L. & Zhang, J. (1996). Conditional Covariance-Based Nonparametric Multidimensionality Assessment. *Applied Psychological Measurement,* 20, 331-354.

Stout, W., Nandakumar, R. & Habing, R. (1996) Analysis Of Latent Dimensionality Of Dichotomously And Polytomously Scored Test Data. *Behaviormetrika,* 23, 1, 37-65.

Stout, W., Froelich, A. G. & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. En A. Boomsma, M.A.J. Duijn, & T.A.B. Snidjers (Eds.), *Essays on item response theory* (357-376). NY: Springer-Verlag.

Walker, C.M., Azen, R. & Schmitt, T. (2006). Statistical Versus Substantive Dimensionality. The Effect of Distributional Differences on Dimensionality Assessment Using DIMTEST. *Educational and Psychological Measurement.* 66, 5, 721-738.

Wilson, D.T., Word, R. & Gibbons, R. (1998). *TESTFACT. Test scoring, Item Statistics and Item Factor Analysis.* Chicago: Scientific Software Internacional, Inc.

Zhang, J. & Stout, W. (1999). The theoretical Detect index of dimensionality and its application to approximate simple structure. *Psychometrika,* 64, 2, 213-249.

Zhang, J. (2007). Conditional Covariance Theory And Detect For Polytomous Items. *Psychometrika,* 72, 1, 69–91.

Fuentes electrónicas

Perkhounkova, Y. & Dunbar, S. B. (1999). Influences of Item Content and Format on the Dimensionality of Tests Combining Multiple-Choice and Open-response ítems: An Application of the Poly-DIMTEST Procedure. Trabajo presentado en Annual Meeting of the American Educational Research Association, Montreal, Canadá, Abril. Consultado el 2 de febrero de 2008, de ERIC www.eric.ed.gov.

Wang, Y & Hocevar, D. (1994). Effects of mathematics Test Content Specificity on Essential Dimensionality in U.S. and Japan Data. Trabajo presentado en *Annual Meeting of the American Educational Research Association*, New Orleans, LA, Abril, 4-8. Consultado el 2 de febrero de 2008, de ERIC www.eric.ed.gov

Dirección de contacto: Luis Lizasoain Hernández. Universidad del País Vasco-Euskal Herriko Unibertsitatea. Avenida de Tolosa, 70. 20018 Donostia - San Sebastián. España. Departamento de Métodos de Investigación y Diagnóstico en Educación. E-mail: luis.lizasoain@ehu.es.

*Los datos de este artículo proceden de la investigación SEC2003-09742, proyecto de I+D financiado por el Ministerio de Ciencia y Tecnología. Este estudio no hubiera sido posible sin la desinteresada colaboración de la Subdirección General de Inspección Educativa de la Comunidad de Madrid, así como de la de cada uno de los inspectores que han participado de modo entusiasta y experto en algunas de las fases críticas de la misma.