Importancia de algunos supuestos «psicométricos» en la calibración y equiparación de las pruebas usadas en la evaluación de los sistemas educativos. Estudio del caso de los «Estándares Nacionales» de México

José Luis Gaviria
Universidad Complutense de Madrid

Covadonga Ruiz de Miguel
Universidad Complutense de Madrid

Resumen

En las evaluaciones de los sistemas educativos nacionales es muy importante el estudio de la evolución de los rendimientos de los alumnos a lo largo del tiempo. Para que sea posible realizar comparaciones, las pruebas de distintos años o de distintos grados en el mismo año deben ser equiparados, utilizando para ello, modelos «psicométricos» adecuados. Estos modelos «psicométricos» asumen ciertos supuestos, que en el caso de no ser tenidos en cuenta, pueden introducir variaciones en los resultados que no corresponden a variaciones de los conocimientos de los alumnos, sino que son un artificio del diseño. En este artículo se explica cómo el diseño de las pruebas y los procedimientos de estimación de las puntuaciones de los sujetos, así como las variaciones en la complejidad dimensional de las pruebas pueden afectar a la equiparación, tanto horizontal como vertical, dando lugar a inconsistencias importantes en los datos de años distintos. Como ejemplo se analiza el caso concreto del programa «Estándares Nacionales» de México.

Palabras clave: evaluación de sistemas educativos, Teoría de Respuesta al Ítem, supuestos en la equiparación, estructura factorial, supuesto de unidimensionalidad.

Abstract: *Importance of Specific 'Psychometric' Assumptions in the Calibration and Comparison of the Tests Used for the Assessment of Education Systems. Mexico's 'National Standards' Case*

The study of the evolution of students' performance through time seems to be especially significant in the assessment of national education systems. In order to make comparisons, tests of different years or different forms of the same year should be put on the same level, using for this purpose suitable 'psychometric' models. These models assume certain assumptions, and in the case of not being taking into account, they may introduce variations in the results which do not correspond to the variation of students' performances, but they turn to be a kind of design's artful devise.

In this article we explain how the design and estimate procedures of individual scores, as well as variations in the dimensional complexity of the tests, may affect comparisons, both horizontally and vertically, thus giving rise to relevant inconsistencies in the results. As an example, this article analyses the specific case of the Mexican programme on 'National Standards'.

Key words: education system assessment, item response theory, comparison assumptions, factorial structure, 'one-dimensionality' assumption.

Introducción

Los datos procedentes de la evaluación de los conocimientos de los alumnos adquieren cada vez más importancia en la determinación de la política educativa. La publicación de los resultados de los programas de evaluación, tanto a nivel nacional como internacional, está dejando de ser un evento puntual para convertirse en una cita periódica que despierta el interés tanto de los expertos como de la población general. Considérese las evaluaciones que realiza el actual Instituto de Evaluación en España, o los programas PISA, TIMSS (Trends in Internacional Mathematics and Science Study) y PIRLS (Progress in International Reading Literacy Study) a nivel internacional.

Un elemento importantísimo de estas evaluaciones es la posibilidad de comparar los resultados de ediciones distintas de la misma evaluación, a fin de juzgar acerca de la evolución del sistema educativo evaluado. Si los resultados no están expresados en la misma métrica, esa comparación es imposible.

La comparativa de los resultados de evaluaciones distintas viene garantizada por el proceso técnico conocido como «equiparación».

En puridad, hay dos procedimientos relacionados, aunque distintos, que a veces se denominan de forma similar. Por una parte tenemos el caso de dos formas distintas del mismo test, construidas para evaluar al mismo tipo de alumnos con el mismo tipo de contenidos y de las que se espera que respondan a las mismas especificaciones y tengan el

mismo comportamiento estadístico. Las puntuaciones así obtenidas hacen referencia a los mismos contenidos, y se espera que sean intercambiables. Por otra parte tenemos el caso de tests distintos construidos para evaluar a alumnos de distintas características con contenidos no exactamente iguales. Se trata de situaciones en las que se desea comparar resultados medios de dos cursos distintos, por lo que el contenido de cada test refleja el contenido específico del curso al que está destinado. Estas puntuaciones no son intercambiables, y en cada curso significan cosas distintas. Aunque este segundo caso suele denominarse «equiparación vertical», (Mislevy, 1992; Mislevy, Jhonson & Muraki, 1992), es más correcto referirse a este tipo de procesos como «Escalamiento para la comparativa», *(Scaling to achieve comparability)* según las normas incluidas en Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985).

En la práctica, la posibilidad de constancia en la escala ha sido posible gracias al desarrollo y utilización de los modelos derivados de la Teoría de Respuesta al Ítem (TRI), (Rasch, 1960; Lord & Novick, 1968; Lord, 1980; Embretson & Reise, 2000; etc.). De hecho, Lord demostró que, en sentido estricto, en el ámbito de la Teoría Clásica de los Tests (TCT) la equiparación sólo era posible cuando los dos tests son, o bien perfectamente fiables, o estrictamente paralelos, es decir, cuando de hecho la equiparación no era necesaria (Lord, 1980, Teorema 13.3.1, p. 198). Naturalmente esto no impidió que en el marco de la TCT se desarrollasen procedimientos que permitiesen llevar a cabo en la práctica los procedimientos de equiparación, tratando de minimizar las distorsiones que la inconsistencia teórica producía.

La Teoría de Respuesta al Ítem, además de resolver las inconsistencias mencionadas, permitió poner en marcha procedimientos antes no contemplados, como la equiparación de grupos no equivalentes o la mencionada «equiparación vertical».

Sin embargo, como señalan Kolen y Brennan, (1995, p. 245), los métodos basados en TRI implican unos supuestos estadísticos y metodológicos mucho más rigurosos. Y la violación de dichos supuestos da lugar a errores sistemáticos, que son mucho más difíciles de controlar o contrarrestar que los errores aleatorios.

Entre los supuestos metodológicos está todo lo relativo al diseño utilizado para aplicar los tests, el número de ítems comunes entre las distintas formas y su distribución en las mismas, sus características técnicas, y sus especificaciones.

Entre los supuestos estadísticos, los más importantes son los que tienen que ver con la estabilidad de la estructura interna del constructo medido en las distintas formas a comparar. En sentido estricto se trata de todo lo relacionado con la dimensionalidad del test y las variaciones en complejidad de esa dimensionalidad en las diferentes versiones. El significado «psicométrico» del supuesto de unidimensionalidad, o,

en su versión más general, del supuesto de independencia local, ha sido estudiado por Hattie (1984), Hambleton y Rovinelli (1986) y Gaviria (1988), entre otros.

Hay que tener en cuenta que los modelos TRI se desarrollaron teniendo en mente constructo muy estables y muy próximos a lo que podríamos denominar características estructurales de los individuos, con una gran estabilidad en su contenido y gran constancia en el tiempo. Sin embargo las aplicaciones más frecuentes ahora mismo son las relacionadas con los resultados de los aprendizajes escolares, que se caracterizan por una mayor ambigüedad en su constructo, con variaciones conceptuales de contenido que se derivan en ocasiones de definiciones políticamente dependientes y con el rasgo peculiar de ser altamente variables en el tiempo por definición. De hecho el objetivo expreso de todo el funcionamiento del sistema educativo es precisamente el logro del aumento constante y, a la mayor velocidad posible, de los aprendizajes escolares.

El establecimiento de un programa de evaluación a gran escala que perdure en el tiempo implica la puesta en marcha de una compleja maquinaria «psicométrica» y metodológica, de cuyo preciso funcionamiento depende la fiabilidad y la utilidad de los resultados obtenidos. Cuando se hayan de tomar medidas de política educativa basadas en los resultados arrojados por las evaluaciones, ha de tenerse la completa certeza de que los cambios y variaciones observadas en las distintas «oleadas» de la evaluación responden a cambios en el sistema, no al propio mecanismo de evaluación.

El incumplimiento de los supuestos «psicométricos» puede tener serias consecuencias que anulan la eficacia de la información obtenida. En este artículo se presenta el estudio de un caso en el que puede comprobarse cómo el no tener en cuenta los mencionados supuestos produce en los datos fenómenos extraños que pueden llevar a interpretar erróneamente la evolución del sistema educativo.

Dado que tanto en España como en América Latina cada vez son más numerosos los programas de evaluación y las instituciones creadas para ponerlos en práctica, y que todavía no hay una acumulación suficiente de experiencia compartida, lo ocurrido con el programa de evaluación de «Estándares Nacionales» de México, tal y como se describe en este artículo, puede resultar de gran utilidad como aviso respecto de las precauciones que han de tomarse en estos casos.

Puesto que en este artículo se analiza un caso de inconsistencia en resultados de evaluación debido a incumplimiento de supuestos «psicométricos» fundamentales, la estructura del mismo no puede responder a la que tradicionalmente se espera de una investigación primaria. Podemos decir que este artículo es un estudio de naturaleza «metaevaluativa». Así que los apartados que siguen se refieren a la estructura de la investigación primaria que aquí se analiza.

Objetivos del programa de evaluación «Estándares Nacionales» de México

El programa de evaluación al que se hace referencia en este artículo se conoció con el nombre de «Estándares Nacionales». La evaluación se planteó con el objetivo de determinar la calidad del desempeño del Sistema Educativo en la República Mexicana, a través de los niveles de logro de los alumnos.

> La medición de los «Estándares Nacionales» se centraba en las habilidades cognitivas asociadas a la lengua y el lenguaje matemático. El objetivo de la evaluación era medir el nivel de desarrollo de habilidades para la comprensión lectora y la resolución de problemas matemáticos. El proceso de concepción del estudio conllevó el análisis del currículo para identificar qué es lo que se espera que los alumnos de educación básica logren como resultado de la escolarización; es decir, cuál es el estándar implícito en el currículo (Secretaría de Educación Pública, México, 2002).

Un aspecto que se comprobó determinante del programa «Estándares Nacionales» tiene que ver con su contenido. Desde un primer momento se decidió que estas pruebas pusiesen más énfasis en la medida de las habilidades cognitivas generales que en los contenidos específicos. Así, aunque las dos áreas de conocimientos, Español y Matemáticas, son sustancialmente importantes como materias instrumentales básicas, se suponía que las destrezas a medir eran transversales.

Diseño

El estudio requirió el diseño de pruebas conforme al modelo llamado *evaluación con referencia a criterio,* para sustentar el establecimiento de niveles de desempeño. Para ello se estructuró un conjunto de cuatro niveles, donde el más alto correspondía al pleno logro del estándar por parte del alumno. Estas pruebas eran elementos fundamentales para el Sistema Nacional de Estándares Educativos: su aplicación cada año desde el curso 1997-98 en Primaria, y desde el curso 1999-00 en Secundaria, proporcionaba información que se estuvo explotando para realizar análisis longitudinales y transversales utilizados para analizar el funcionamiento del sistema.

La principal característica de este estudio era su naturaleza longitudinal, por lo que se hacía especial hincapié en la evolución del rendimiento a través de los años. Esto suponía la comparativa de los resultados tanto entre los grados como entre los años, distintos en un mismo grado. Como consecuencia era necesario posibilitar tanto la equiparación vertical, (comparativa entre grados), como horizontal (comparación entre grupos distintos de los mismos grados). Para ello se hacía necesaria la utilización de un modelo «psicométrico» que posibilitase el cumplimento de estas condiciones. Se optó, debido a ello, por utilizar el modelo de Rasch.

Dado el carácter longitudinal del estudio, la muestra de cada grado que se seleccionó el primer año se constituyó en una cohorte cuyo desempeño fue monitorizado año tras año. Es decir, la muestra de alumnos que en 1998 estaba en primer grado fue evaluada otra vez al año siguiente cuando estaba en segundo grado, y así sucesivamente. Para mantener la representatividad de la muestra respecto a la población nacional de referencia, se introducían cada año las correcciones necesarias.

En Martínez (2000), Martínez y Schmelks (2000) y Martínez-Rizo (2004) se incluye una descripción muy detallada del programa. El último autor señala algunas objeciones a su diseño.

Muestra

La composición de la muestra a la cual se aplicaba las pruebas debía permitir hacer comparaciones en varios niveles, según el grado de desagregación de interés. Por ejemplo, las comparaciones podían ser del total de la población nacional a lo largo de diferentes años, o en un momento determinado, o referirse a las entidades federativas.

La muestra se extrajo con los estratos siguientes: en Primaria en escuelas públicas urbanas, particulares, públicas rurales, educación indígena y cursos comunitarios; y en Secundaria, se analizaron los estratos de Secundaria general, técnica y telesecundaria.

Las Tablas I y II presentan las muestras de alumnos participantes por grado evaluado en cada una de las evaluaciones de datos realizadas hasta 2002.

TABLA I. Escuelas y alumnos evaluados en Estándares Nacionales en Primaria

		Grados						
Ciclo escolar	**Escuelas**	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**Totales**
1997-98	3.310	44.218	43.398	48.866	45.214	43.044	32.255	256.995
1998-99	3.193			49.136		46.252	32.622	128.010
1999-00	3.406		48.891		47.934		45.676	142.501
2000-0	3.367			49.068		46.574	39.562	135.204

TABLA II. Escuelas y alumnos evaluados en Estándares Naciones en Secundaria

		Grados			
Ciclo escolar	**Escuelas**	**1º**	**2º**	**3º**	**Totales**
1999-00	1.144	37.695	37.325	36.584	111.604
2000-0	1.193	39.354	38.886		78.240

Metodología

Como consecuencia de que se pretendía evaluar habilidades cognitivas consideradas como transversales, el diseño de equiparación que se adoptó se preocupaba exclusivamente porque hubiese una cadena ininterrumpida de ítems entre todos los grados y todos los años, sin atender al área de conocimiento a que estos ítems pertenecían. Ocurrió entonces que algunos años los ítems comunes pertenecían a una de las áreas, Matemáticas, y en otras ocasiones a la otra, el área de Español, sin que se garantizase la continuidad sin interrupción en cada cadena (Español y Matemáticas) por separado. Tampoco todos los años y grados contenían la misma proporción de ítems de Matemáticas y de Español. Como veremos más adelante, este hecho tuvo importantes consecuencias prácticas.

A esta concepción de la prueba se unió la práctica de proporcionar, como resultados de la evaluación, tres puntuaciones distintas basadas en una calibración única de los ítems. Así, se proporcionaba una nota general, otra para Matemáticas y otra para Español. Sin embargo no se llevaban a cabo tres estimaciones independientes, lo que por el diseño elegido hubiera sido imposible, sino una global, que incluía a todos los ítems, cada vez estableciendo como fijas las medidas de los ítems usados como ancla.

Resultados y discusión

Transcurridos varios años en los que se hizo el oportuno levantamiento de datos, se procedió al análisis de los mismos y a estudiar la evolución en el tiempo de los aprendizajes de los alumnos. Es en ese momento cuando se hacen patentes algunas posibles inconsistencias respecto de las que conviene dilucidar si reflejan algún fenómeno notable en el sistema educativo, o si se trata de un artefacto del proceso de medición.

Analizamos en primer lugar lo que ocurre con los datos obtenidos en Matemáticas (véase Tabla III. Resultados de «Estándares Nacionales» de Matemáticas). Para evitar ser prolijos, debemos señalar que dado el gran tamaño de la muestra utilizada, prácticamente todas las diferencias observadas son estadísticamente significativas.

TABLA III. Resultados de «Estándares Nacionales» de Matemáticas

	Matemáticas								
	Primaria						Secundaria		
Grado	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**1º**	**2º**	**3º**
1998	414,2	436,8	423,9	434,0	464,9	480,0			
1999			431,7		467,2	490,7			
2000		422,5		472,9		499,8	515,1	511,4	526,1
2001			424,2		480,7	511,2	508,3	520,9	540,1
2002		378,0		428,5		493,3	513,8	525,1	

La idea básica que gobierna este análisis es que un sistema educativo es un subsistema social con una gran inercia funcional. Por tanto los cambios que se produzcan en el mismo han de ser paulatinos, y no constituyen respuestas inmediatas a decisiones políticas puntuales. De este modo los efectos de cualquier cambio político-organizativo han de propagarse progresivamente por el sistema. Cambios drásticos e inesperados en las tendencias son síntoma de posibles problemas en la metodología de la evaluación.

Con esta perspectiva, cuando analizamos los resultados vemos varias cosas que merecen ser analizadas con detalle. Así, en algunos casos tenemos grados superiores que obtienen puntuaciones medias menores que grados inferiores. Por ejemplo:

- En 1998 la cohorte de 2º de Primaria obtiene 436,8 puntos, mientras que la de 3º de Primaria obtiene 12,9 puntos menos (423,9) y la de 4º de Primaria obtiene también un puntaje inferior a la de 2º (434,0 puntos).

- En el año 2001, los alumnos de 1º de Secundaria, a quienes se les supone un rendimiento superior a los alumnos de 6º de Primaria, obtienen 2,9 puntos menos que éstos (511,2 frente a 508,3).

En otras ocasiones una misma cohorte de alumnos obtiene puntuaciones inferiores cuando se encuentra en grados superiores frente a los que obtenía cuando se encontraba en grados inferiores:

- La cohorte de 2º de Primaria en 1998 obtiene 436,8 puntos, y en contra de lo que cabría esperar, esa misma cohorte, en 3º de Primaria, en 1999, obtiene 5,1 puntos menos (obtiene 431,7 puntos). Ha empeorado. Sin embargo, ese mismo grupo, cuando pasa a 4º de Primaria, en el año 2000, aumenta 41,2 puntos, obteniendo 472,9 puntos. Cuando pasa a 5º de Primaria, en el año 2000, el aumento es sólo de 7,8 puntos, llegando hasta 480,7. En 6º de Primaria, esa cohorte, aumenta 12,6 puntos, obteniendo 493,3 puntos.

En otros momentos los incrementos de un grado al siguiente no son homogéneos entre cohortes ni tampoco en el interior de una misma cohorte:

- Cuando el grupo que en el año 2000 está en 2º de Primaria pasa a 3º de Primaria en el 2001, la diferencia en sus puntuaciones es de 1,7 puntos (pasa de 422,5 a 424,2). Cuando ese grupo pasa a 3º de Primaria, en 2002, el aumento es de 4,3 puntos (de 424,2 a 428,5 puntos). Se trata pues de incrementos muy modestos.
- Sin embargo, cuando la cohorte que en 1999 está en 3º de Primaria pasa a 4º (año 2000), el aumento es nada menos que de 41,2 puntos (de 431,7 a 472,9).

En Secundaria, en el año 2000 se da una situación, cuando menos, curiosa. El rendimiento de los alumnos de 1º curso es 3,7 puntos superior al de el 2º curso (515,1 y 511,4 respectivamente). Sin embargo cuando esas dos cohortes pasan al grado superior, 2º y 3º de Secundaria en 2001 respectivamente, la diferencia se invierte y pasa a ser de 19 puntos a favor de los alumnos del curso superior. También ocurre que los que en 2001 se encontraban en 1º de Secundaria, obtuvieron una media inferior a los del año anterior, pero al pasar al grado siguiente mejoran el rendimiento de aquéllos.

La cohorte de alumnos de 1º de Secundaria en el año 2000 obtenía un rendimiento de 515,1 puntos, y la que se toma en 2001 baja en 2,9 puntos, hasta quedarse en 508,3 puntos.

Si se observa la evolución del rendimiento de los grupos de 5º de Primaria de los años 1998 y 1999 hasta que llegan a 2º de Secundaria, puede verse que aumenta cada año. Sin embargo, si vamos comparando las cohortes año a año, puede verse que para 5º y 6º de Primaria y 2º de Secundaria, la segunda cohorte obtiene un rendimiento superior a la de la primera. Pero esta tendencia se invierte para el caso de 1º de Secundaria.

La ganancia de los alumnos desde 3º de Primaria (cohorte de 1998), que obtienen 423,9 puntos hasta 6º de Primaria, que obtienen 511,2, es de 87,3 puntos. La ganancia de los alumnos desde 2º de Primaria (cohorte de 1998), que obtienen 436,8 puntos hasta 6º de Primaria, que obtienen 493,3 es de 56,5 puntos. A pesar de tener un grado más en el primer caso, la ganancia ha sido 30,8 puntos menos en el segundo. Sin embargo, esta misma cohorte, que había aumentado 87,3 puntos de 3º a 6º de Primaria, sólo es capaz de aumentar 2,6 puntos al pasar a 1º de Secundaria (513,8 puntos)

En 6º de Primaria la cohorte de 1999 es 20,5 puntos mejor que la de 2001 (obtienen 490,7 y 511,2 puntos respectivamente). Sin embargo, estas mismas cohortes, al compararlas en 1º de Secundaria invierten la diferencia. Cuando la cohorte de 1999 pasa a Secundaria obtiene 515,1 puntos, y cuando lo hace la de 2001, obtiene sólo 513,8, o lo que es lo mismo 1,3 puntos menos.

También en la asignatura de Español aparecen varias inconsistencias que merece la pena reseñar.

TABLA IV. Resultados de «Estándares Nacionales» de Español

	Español								
	Primaria						Secundaria		
Grado	**1º**	**2º**	**3º**	**4º**	**5º**	**6º**	**1º**	**2º**	**3º**
1998	407,0	419,3	417,4	437,2	456,8	474,5			
1999			432,1		466,7	488,4			
2000		427,5		475,0		500,9	507,9	506,5	523,9
2001			425,9		479,7	505,0	500,9	535,8	
2002		377,5		437,6		501,8	504,1	533,3	563,1

Se dan casos en los que rendimientos medios de un grado son menores que los de grados inferiores. En efecto, en contra de lo cabría esperar, en el año 2000 la cohorte de 2º de Secundaria obtiene un rendimiento inferior en 1,4 puntos respecto de la de 1º de Secundaria. La cohorte que en 2000 está en 3º de Secundaria, obtiene un rendimiento de 523,9 puntos, mientras que la de 2º de Secundaria del año siguiente (2001) obtiene, a pesar de ser un curso inferior, 11,9 puntos más.

También encontramos inconsistencias cuando comparamos los incrementos de resultados de los distintos grados. Por ejemplo, la cohorte de 2001 para 1º de Secundaria es 7 puntos peor que la de 2000. Sin embargo los alumnos de 2º de Secundaria tienen una media 29 puntos mayor que los del año anterior.

Tampoco hay constancia en los comportamientos de las cohortes. Así, la de 6º de Primaria del año 2000 no obtiene ninguna mejoría de rendimiento al pasar a 1º de Secundaria, su rendimiento es de 500,9 puntos en ambos casos. Sin embargo en el curso siguiente estos alumnos mejoran en 32,4 puntos, llegando hasta los 533,3 puntos.

En lo que sigue trataremos de explicar cómo la existencia de estas inconsistencias puede deberse a dos tipos de problemas: unos relacionados con el procedimiento utilizado para obtener estimaciones de puntuaciones de Matemáticas y Español, y otros relacionados con las variaciones en la complejidad de las estructuras factoriales de las pruebas utilizadas en los distintos años.

Propiedades del modelo relevantes en la evaluación analizada

Como hemos visto, un objetivo fundamental de este programa de evaluación era garantizar la comparativa de los resultados. Ésta es la razón por la que se decidió utilizar el modelo de Rasch (1960).

Según este modelo la respuesta a un ítem depende sólo de la competencia del sujeto (θ_j) y de la dificultad del ítem (b_i). La curva característica del ítem (CCI) tiene la expresión siguiente:

$$P(u_{ij} = 1 \mid \theta_j) = \frac{1}{1 + \exp(-(\theta_j - b_i)}$$

Tanto la capacidad del sujeto, θ_j, como la dificultad del ítem, b_i, están en la misma escala, es decir, son comparables e intercambiables. Además, la probabilidad de la respuesta correcta está sólo en función de θ_j y de b_i, independientemente de qué o cuántos sujetos tienen la capacidad θ_j. Por otro lado, la capacidad del sujeto es invariante respecto del subconjunto específico de ítems que responda. Todas estas propiedades dependen del cumplimento del supuesto de unidimensionalidad, o de modo más general, del de independencia local.

Según el supuesto de unidimensionalidad, la probabilidad de responder correctamente a un ítem depende sólo de θ. La TRI asume en su formulación que los ítems destinados a medir la variable θ constituyen una sola dimensión.

El supuesto de unidimensionalidad es un caso particular de otro supuesto, el Supuesto de Independencia Local (SIL). Según éste, las respuestas de una persona a los ítems a que se somete son estadísticamente independientes entre sí. Es decir, dado un valor de θ_j, el rendimiento en un ítem no viene determinado por los resultados obtenidos en otros, sino sólo por el rasgo que el ítem mide.

Dicho de otro modo, la independencia local puede expresarse diciendo que la probabilidad de que un sujeto acierte «n» ítems, dada su capacidad $\prod j$, es igual al producto de las probabilidades de acertar cada uno de ellos.

$$p(U_{1j}, U_{2j}, ..., U_{pj} \mid \theta_j) = p(U_{1j} \mid \theta_j) p(U_{2j} \mid \theta_j) ... p(U_{pj} \mid \theta_j)$$

Esta es la función de verosimilitud. Así, el supuesto de unidimensionalidad nos permite definir la función de verosimilitud y obtener estimaciones de los parámetros.

El modelo cumple varias condiciones de invariabilidad: a) las puntuaciones de los sujetos son invariantes respecto del subconjunto de ítems utilizado; y b) los parámetros de los ítems son invariantes respecto de la muestra empleada. Esta invariabilidad, tan importante para la equiparación, depende del cumplimiento del SIL.

Pueden entenderse las implicaciones prácticas de estas propiedades sin entrar en detalles matemáticos.

En las Ilustraciones I, II y III aparece la representación gráfica de diferentes ejecuciones de un sujeto en diferentes tests que miden la misma variable θ. X representa el número de aciertos que consigue en cada test. En el gráfico aparece también el porcentaje de aciertos con respecto al test completo, y como puede verse, independientemente de esto, el valor de θ para el sujeto siempre es el mismo.

El parámetro θ no depende de que utilicemos un test u otro. Por ejemplo, comprobemos qué ocurriría si a un sujeto (o grupo de sujetos) con $\theta = 0{,}5$ se le aplicasen tres tests, el primero con 17 ítems, el segundo con 9, y el tercero con 10, ordenados según b (índice de dificultad), y que midiésemos sin error[1].

La recta representa el continuo que queremos medir. Cada uno de los ítems está representado por una marca sobre la recta. La posición de la marca correspondiente estará más a la izquierda cuanto más fácil sea el ítem, y más a la derecha cuanto más difícil. La posición del sujeto con capacidad θ viene señalada por la marca que aparece destacada. También la posición del sujeto en la recta está indicando cuál es su capacidad. En

(1) Es en este contexto cuando asumimos que lo que queremos decir, cuando indicamos que «medimos sin error», es que un sujeto siempre contestará correctamente a aquellos ítems cuyo índice de dificultad es inferior a su capacidad y; siempre contestará incorrectamente a aquellos ítems cuya dificultad sea superior a esa misma capacidad.

la Ilustración I vemos que la capacidad de este sujeto es mayor que la dificultad de los ocho primeros ítems, e inferior que la de los nueve siguientes. Si como hemos asumido, medimos sin error, este sujeto responderá correctamente a los primeros ocho ítems, ya que para él son fáciles. Sin embargo contestará incorrectamente a los nueve siguientes, ya que para él son difíciles. (Es decir, la dificultad de los ítems es mayor que su capacidad). Comprobamos que en estas condiciones, tanto la puntuación observada clásica (número de respuestas correctas), como el porcentaje de respuestas correctas, no es invariante, es decir, depende de cuántos ítems contenga el test que conteste el individuo.

Vemos que en el primero de ellos acertaría el 47% de los ítems (8) (véase Ilustración I), en el segundo el 89% (8 ítems) (véase Ilustración II), y en el tercero sólo el 10% (1 ítem) (véase Ilustración III). Sin embargo, el valor de θ permanece invariable en los tres casos.

ILUSTRACIÓN I.

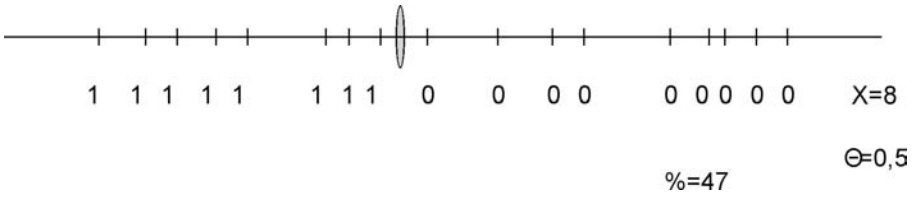


ILUSTRACIÓN II.

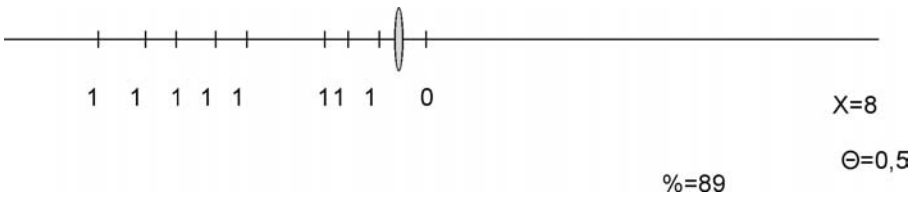
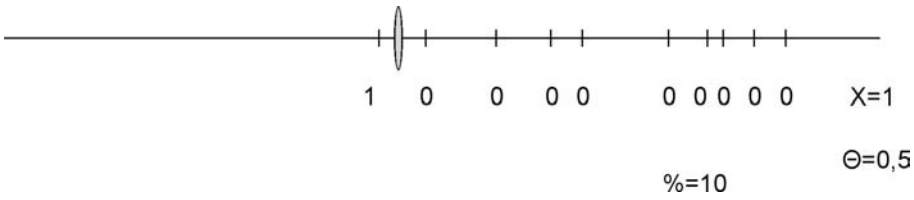


ILUSTRACIÓN III.



Basándonos en estas propiedades podemos afirmar que las inconsistencias observadas en los datos pueden producirse de al menos dos maneras distintas, relacionadas con las características específicas del diseño del programa de evaluación que estamos analizando.

Es posible, en primer lugar, que al elegir subconjuntos distintos de ítems para hacer la estimación de las puntuaciones, Matemáticas por un lado y Español por otro, la distinta distribución de estos dos conjuntos en las pruebas de cada grado den lugar a desplazamientos de las medias estimadas que no responden a verdaderas diferencias en los rendimientos de los alumnos.

Como ya hemos señalado, la otra posibilidad es que no se cumpla el supuesto de unidimensionalidad, y que las pruebas de los distintos grados tengan distintas estructuras factoriales, con lo que las diferencias estimadas no reflejan cambios en los rendimientos de los alumnos sino cambios en los instrumentos de medida.

En los puntos siguientes se analizan estas dos posibilidades.

El procedimiento de estimación como posible origen de las inconsistencias

Si asumimos que se trata de una sola habilidad la que queremos medir y calibramos todos los ítems en la misma escala, las puntuaciones de los sujetos serán invariantes respecto del subconjunto específico de ítems que utilicemos para su estimación.

Si se cumpliesen a la perfección todos los supuestos, especialmente el de unidimensionalidad, la estimación de las thetas (θ) obtenidas utilizando todos los ítems, sólo los ítems de Español o sólo los ítems de Matemáticas, tendrían la misma esperanza matemática.

Las puntuaciones por separado de Español y de Matemáticas, no son estimaciones de dos cosas distintas, sino estimaciones con mayor incertidumbre de la misma capacidad latente. No tendría sentido proporcionar puntuaciones distintas de la puntuación global. No son cosas distintas.

Sin embargo, si al calibrar los ítems se asumió que se trataba de una escala común, como hemos explicado, la esperanza del estimador de la puntuación de cada sujeto es la misma, independientemente de si consideraron todos los ítems, sólo los de Matemáticas, o sólo los de Español. ¿Por qué entonces tenemos tres puntuaciones distintas para cada sujeto?[2]. En las Ilustraciones IV y V podemos ver la respuesta.

[2] Aquí sólo se han analizado los resultados de Español y Matemáticas, dejando fuera de nuestro análisis la escala global.

En la Ilustración IV representamos un test compuesto de ítems de Matemáticas y de Español. Si asumimos el modelo de (1), entonces postulamos implícitamente que todos los ítems están situados sobre un continuo como el representado en la Ilustración IV. La flecha indica la posición de la puntuación latente $\prod$ para un sujeto determinado. Asumimos otra vez el supuesto de la medición sin error. Si el sujeto responde a todos los ítems, la estimación de θ que obtendremos para él vendrá dada por la marca de color, cuya posición, en este caso, viene a coincidir con la de la flecha.

Así que, si sólo utilizamos los ítems de Español, (véase Ilustración V) la mejor estimación posible viene dada por la marca de color. La puntuación que corresponde con la capacidad latente del sujeto (la flecha) no ha cambiado, pero la estimación ahora difiere de la anterior. Naturalmente esto no quiere decir que estemos tratando de dos magnitudes distintas, sino de dos estimaciones obtenidas con distinta precisión, de la misma capacidad latente θ.

Una primera consecuencia para las pruebas de «Estándares Nacionales» es que las estimaciones de las puntuaciones de Español y Matemáticas no son independientes. En consecuencia la no coincidencia de esos valores es, en sí misma, una primera indicación del nivel de incertidumbre que se introduce al hacer las estimaciones de las medias basándose en uno u otro conjunto de ítems. En segundo lugar, las inconsistencias entre cursos pueden deberse a la distinta distribución de las proporciones de ítems de Español y Matemáticas en cada grado. En efecto, incluso cumpliéndose el SIL, el nivel de incertidumbre producido por la distinta distribución de los ítems de cada materia difumina las verdaderas diferencias de un grado al siguiente.

ILUSTRACIÓN IV. Estimación con todos los ítems

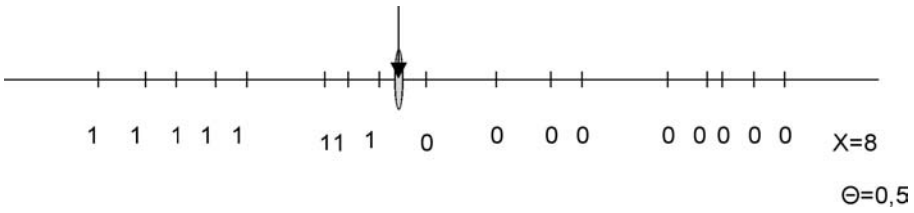
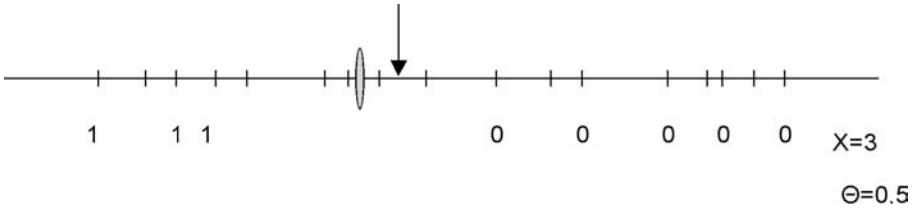


ILUSTRACIÓN V. Estimación con los ítems de Español



La distinta complejidad de estructura factorial como posible origen de las inconsistencias

¿Cómo afecta el supuesto de unidimensionalidad al contenido de las pruebas de evaluación? Podemos decir que si queremos medir contenidos que desde un punto de vista escolar sean relevantes, es muy difícil que podamos satisfacer las condiciones tradicionales de unidimensionalidad. Pero se puede adoptar lo que podemos llamar una «unidimensionalidad débil».

Según esto, en la práctica, la condición esencial para que podamos hablar de unidimensionalidad es que los sujetos puedan ser ordenados unívocamente a partir de sus puntuaciones en la magnitud medida. Así, una condición necesaria, aunque no suficiente, para que dos pruebas distintas definan una sola dimensión es que cada una de ellas por separado proporcione la misma ordenación de los sujetos. Para que esto sea posible es necesario que las dos pruebas tengan la misma estructura factorial, con proporciones idénticas de ítems de todas las dimensiones y distribuciones similares de las dificultades relativas en cada dimensión.

Cuando dos pruebas no mantienen la misma estructura factorial, se producen consecuencias en la equiparación, y es que se producirán variaciones aparentes en las medias de las puntuaciones estimadas, sin que necesariamente hayan cambiado los rendimientos de los alumnos.

A efectos ilustrativos vamos a asumir que las pruebas de «Estándares Nacionales» estuviesen compuestas por sólo dos factores asociados a los contenidos básicos de Español y Matemáticas. Como luego se verá, la realidad es bastante más compleja, pero todo lo que aquí se presenta es generalizable a situaciones de mayor complejidad factorial.

En la Ilustración VI aparecen representadas las puntuaciones de dos sujetos en los componentes de Matemáticas y Español de las pruebas. Como podemos comprobar, las capacidades de estos sujetos son simétricas. La puntuación que el primero tiene en Español, el segundo la tiene en Matemáticas, y viceversa.

En la Ilustración VII, se asume que en la prueba de estándares tienen el mismo peso los ítems de Matemáticas y Español. En estas condiciones los dos sujetos tienen la misma puntuación en «Estándares Nacionales».

Sin embargo, en las Ilustraciones VII y VIII se asume que el peso de Matemáticas es mayor en la composición de la prueba, primer caso, o que es mayor el peso de Español, segundo caso. Las puntuaciones estimadas de los sujetos son en esos dos casos diferentes. El sujeto con mayor capacidad matemática (punto 1) tiene mayor

puntuación en la segunda versión, mientras que el que tiene mayor capacidad lingüística (punto 2) obtiene mayor puntuación en la tercera versión. Como se comprueba, los cambios en las puntuaciones de «Estándares Nacionales» reflejan cambios en la prueba, no cambios en los sujetos.

ILUSTRACIÓN VI. Puntuaciones simétricas de dos sujetos en dos factores. Peso equivalente de los factores en la prueba de «Estándares Nacionales»

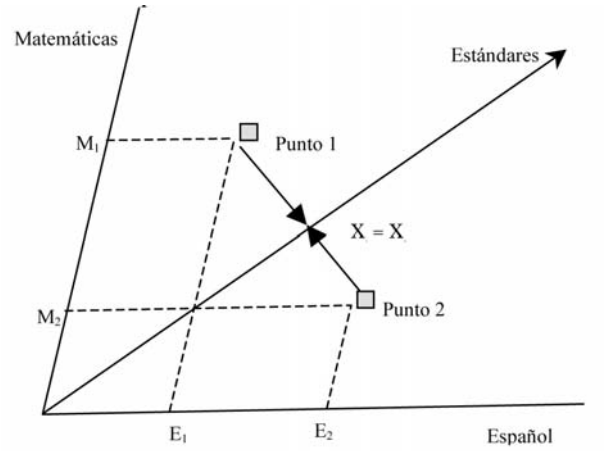


ILUSTRACIÓN VII. Puntuaciones simétricas de dos sujetos en dos factores. Mayor peso del factor Matemáticas en la Prueba de «Estándares Nacionales»

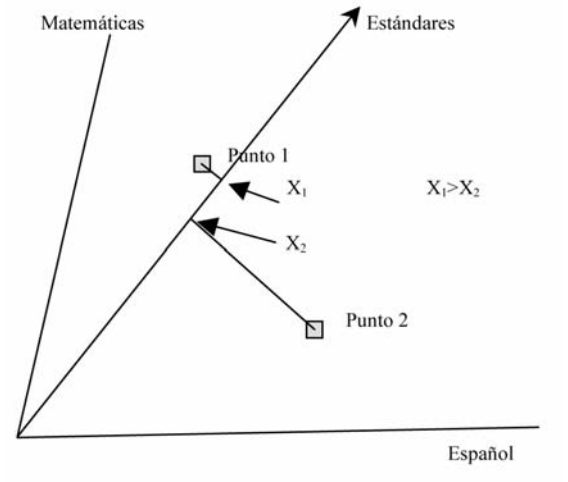
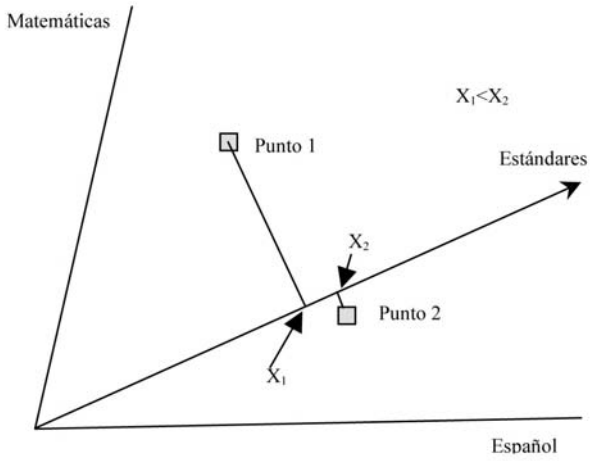


ILUSTRACIÓN VIII. Puntuaciones simétricas de dos sujetos en dos factores. Mayor peso del factor Español en la Prueba de «Estándares Nacionales»



Respecto de la asunción de unidimensionalidad en las pruebas de «Estándares Nacionales»

La elección de estas habilidades cognitivas como objetivo de la evaluación no deja de presentar algunos problemas conceptuales. Ciertamente podemos afirmar que no hay pedagogo moderno que no considere que en el dominio cognitivo, el objetivo de la educación debe consistir más en «enseñar a pensar» que en «enseñar datos». Este «aprender a pensar» tendría que ver con la utilización de procesos generales de razonamiento y resolución de problemas transferibles a distintas áreas sustantivas. Si el verdadero núcleo de las pruebas estuviese constituido, no por el contenido material de las dos áreas mencionadas, sino por las habilidades comunes a varias materias, serían éstas las que proporcionarían el sustento conceptual de la unidimensionalidad, requisito fundamental para la equiparación.

Pero desde un punto de vista metodológico es discutible que sean las habilidades definidas como una suerte de destrezas «metacognitivas» que serían independientes de la materia enseñada, el objetivo o el producto del trabajo escolar. No se discute aquí la importancia de estas destrezas cognitivas. Lo que se discute es que ese sea el producto inmediato del trabajo escolar.

Supongamos que esas «habilidades» se enseñan en la escuela. Si es así, se enseñan en cada materia. Si lo enseñado en una materia es transferido a otras, y eso justifica la unidi-

mensionalidad, entonces sería superflua la enseñanza de todas las demás materias. Bastaría con una sola enseñanza y, por ende, con la medida de un solo ámbito de contenido. Si por el contrario las habilidades deben aprenderse en combinación con distintos contenidos, entonces es posible que se haya aprendido una habilidad con las Matemáticas pero no con la Lengua Española. Por tanto no deberíamos tener una sola medida para cada sujeto, sino dos. Pero eso parece que sería contradictorio con el concepto de «habilidad».

Cabe también la posibilidad de que las habilidades a que se refieren supuestamente las pruebas de «Estándares Nacionales» sean en realidad la combinación de unos contenidos con unas capacidades innatas de los sujetos. La justificación de la prueba estaría entonces en que mide, no el resultado de la integración de esos dos elementos, sino el que es común a las materias, es decir, la capacidad transversal. Si esto es así, estas pruebas están midiendo en realidad el desarrollo madurativo o la madurez intelectual de los sujetos, y no el resultado del trabajo escolar.

¿Qué sentido tiene medir una característica estructural de los sujetos, es decir, una característica con un fuerte componente hereditario, y por tanto, poco susceptible de variación, para valorar el trabajo en la escuela?

Pero es posible comprobar empíricamente hasta qué punto esa supuesta unidimensionalidad está presente en los datos.

De entre los diferentes métodos para comprobar la unidimensionalidad de los ítems, recopilados por Hattie, 1984, 1985; Hambleton y Rovinelli, 1986; o Stout, 1987, el análisis factorial sigue siendo el más utilizado (Muñiz, 1990).

Son pocas las ocasiones en las que se va a encontrar una unidimensionalidad perfecta (que un sólo factor explique el 100% de la varianza), por lo que la unidimensionalidad se convierte en un asunto de grado.

En la Tabla III tenemos la varianza explicada por el primer factor correspondiente a los datos de los años 1998 a 2002 para cada uno de los grados.

TABLA V. Primer autovalor y porcentaje de varianza explicado en las pruebas de «Estándares Nacionales» de 1998 a 2002

Año 1998

Grado	Primer autovalor	Número de variables	% de Varianza explicada por el primer factor
1 Prim	13,258	55	24,11
2 Prim	11,980	48	24,96
3 Prim	8,340	49	17,02
4 Prim	7,980	53	15,06
5 Prim	6,748	55	12,27
6 Prim	5,930	55	10,78

Año 1999

Grado	Primer autovalor	Número de variables	% de Varianza explicada por el primer factor
3 Prim	13,687	90	15,21
5 Prim	47,970	94	51,03
6 Prim	12,249	93	13,17

Año 2000

Grado	Primer autovalor	Número de variables	% de Varianza explicada por el primer factor
2 Prim	15,424	74	20,84
4 Prim	17,286	87	19,87
6 Prim	15,103	91	16,60
1 Sec	14,126	114	12,39
2 Sec	15,894	113	14,07
3 Sec	12,731	106	12,01

Año 2001

Grado	Primer autovalor	Número de variables	% de Varianza explicada por el primer factor
3 Prim	12,987	86	15,10
5 Prim	14,084	105	13,41
6 Prim	15,947	103	15,48
1 Sec	15,875	125	12,70
2 Sec	18,271	125	14,62

Año 2002

Grado	Primer autovalor	Número de variables	% de Varianza explicada por el primer factor
2 Prim	13,461	76	17,71
4 Prim	12,899	94	13,72
6 Prim	17,407	107	16,27
1 Sec	14,042	119	11,80
2 Sec	15,867	119	13,33
3 Sec	17,768	119	14,93

Media =	16,86
Desviación típica =	7,83

Como puede verse, salvo en el caso de 5º de Primaria del año 1999, en el que la varianza explicada es 51,03%, el resto de los valores son muy bajos, no superando el 25% en ningún caso.

Para hacernos una idea de la importancia del problema, en la Ilustración X tenemos representada la «comunalidad» de cada uno de los ítems de una de las pruebas, asumiendo cuatro factores. Los ítems han sido ordenados en orden creciente de «comunalidad» de izquierda a derecha. De 125 ítems que componen esa prueba concreta, alrededor de 98 presentan una «comunalidad» menor del 30%, y sólo 3 ítems superan el 50%.

ILUSTRACIÓN IX. Porcentaje de varianza explicada en cada grado cada año por el primer factor

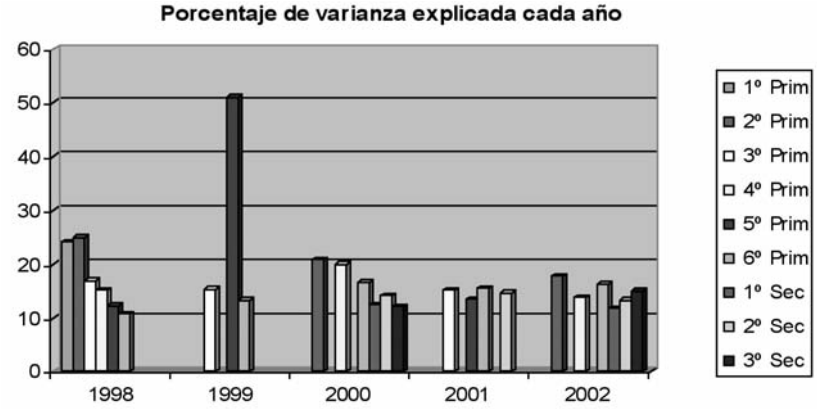
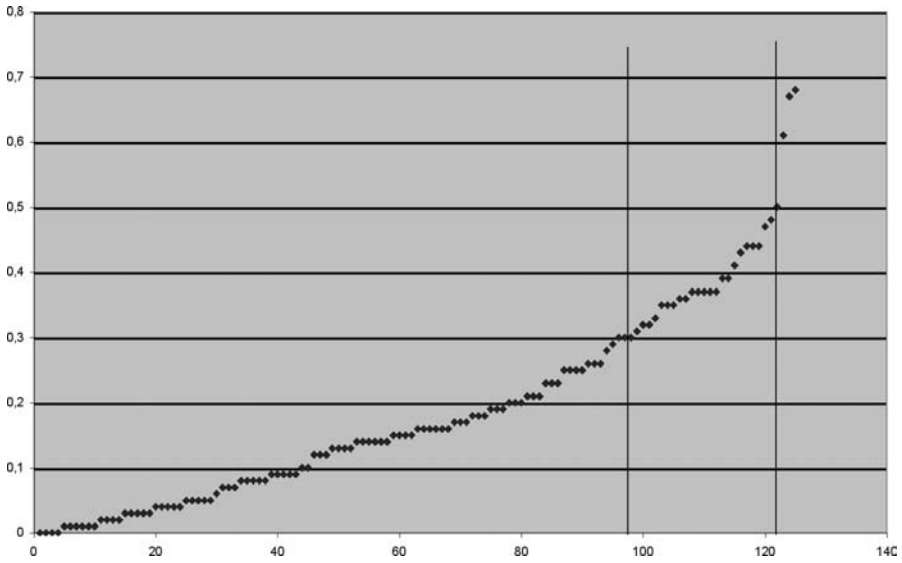


ILUSTRACIÓN X. Comunalidades con cuatro factores. 2º Secundaria 2001



Posibilidad y conveniencia de la unidimensionalidad en la evaluación de sistemas educativos

Ante estos resultados surgen naturalmente dudas respecto a la propia metodología que subyace a la evaluación. ¿Es posible cumplir el supuesto de unidimensionalidad cuando se está tratando de evaluar constructos complejos como el aprendizaje producto del trabajo escolar? ¿No es un supuesto excesivamente restrictivo? ¿No hace este supuesto inviable la utilización de la TRI?

En este punto debemos señalar que la existencia de unidimensionalidad no implica necesariamente que el rasgo unidimensional que se pretende medir deba ser necesariamente simple. La confusión entre singularidad conceptual y unidimensionalidad ha llevado en ocasiones a errores en la interpretación de la naturaleza de algunas variables. Citando textualmente a Lumsden (1976):

> Es posible que la preocupación acerca de la dificultad de la construcción de test unidimensionales surja de una confusión entre unidimensionalidad y sin-

gularidad teórica. Un test unidimensional tiene un sólo atributo, pero el atributo puede ser complejo. Consideremos el siguiente ítem:

1 2 4 7 11 — — —

Un test con ítems de este tipo es unidimensional y describe destreza del sujeto al responder a ítems como ese (…). La destreza es compleja, implicando capacidades numéricas y de razonamiento y probablemente también otras. No refleja un constructo o atributo teórico sencillo en la terminología de Cronbach y Meehl. Puede entenderse como un compuesto de constructos como elementos. La construcción de tests teóricamente singulares es probablemente imposible (p. 267).

Esa imposibilidad tiene que ver con la amplitud del rasgo a medir. Cuando decimos que un constructo es complejo, que tiene varias dimensiones, ¿qué entendemos por las dimensiones fundamentales de un constructo?, ¿cómo aparecen?, ¿cómo se caracterizan?, ¿qué relaciones hay entre esas dimensiones fundamentales para determinar el constructo?

El análisis factorial, una técnica que ha sido utilizada en el análisis de los ítems, nos proporciona un marco conceptual ideal para revisar algunas de estas cuestiones.

Cuando se analiza de forma factorial un test, los factores principales se han utilizado para seleccionar subconjuntos homogéneos de ítems y construir con ellos escalas unidimensionales. Se considera entonces que se ha descubierto la estructura interna del test, ya que han aparecido las dimensiones fundamentales. Nada más lejos de la realidad.

En efecto, supongamos que, en un ejemplo clásico utilizado para explicar el análisis factorial, estamos tratando de descubrir la naturaleza de la variable volumen. Para ello contamos con un gran número de recipientes paralelepípedos de distintos tamaños y proporciones. Sobre esos recipientes realizamos varias medidas, y luego analizamos de forma factorial los resultados obtenidos. Si el tipo de medidas que hemos obtenido se refieren a longitudes, alturas, diagonales, etc., aparecerán, al menos, tres factores independientes, que de alguna forma nos dicen en qué espacio está el constructo volumen. Pero sin embargo no sabemos todavía nada acerca del mismo. Es más, el análisis factorial no nos sirve para determinar cómo son las relaciones entre esas dimensiones, ni cómo a partir de la información que nos proporcionan podemos obtener información acerca del volumen. Es preciso que antes sepamos que el volumen es el producto de las tres dimensiones fundamentales. Pero ese conocimiento no

nos vendrá proporcionado por el análisis factorial; es *a priori.* Por otra parte, el constructo volumen presenta una estructura compleja en la que están implicados los tres factores que subyacen a las medidas que hemos tomado. ¡Pero con otras medidas distintas, el volumen aparecería como una variable simple! Por ejemplo, supongamos que nuestras medidas hubiesen consistido en el peso de cada recipiente cuando se llena de distintas sustancias, agua, mercurio, grava, arena, aire, etc. Tendríamos tantas variables como sustancias. Pero nadie puede dudar de que el resultado del análisis factorial de tales datos se ajustaría perfectamente a un modelo de un sólo factor común.

El investigador que hubiese llevado a cabo estos análisis habría llegado a la conclusión de que se trata de una variable teóricamente simple. Está claro que la complejidad estructural de un test, o de un constructo, no viene sólo determinada por la técnica utilizada para analizarlos sino también por el conocimiento previo que tengamos acerca del tema.

En el segundo caso, el análisis factorial nos serviría para conocer la «naturaleza» del constructo, pero no las dimensiones fundamentales del mismo, y en el primero lo contrario.

Esta última cuestión nos lleva a plantearnos precisamente la forma en que se utilizan, para la interpretación, los resultados del análisis factorial. Por lo general, rara vez se trata de buscar una entidad conceptual que trate de explicar simultáneamente todas las conductas que quedan agrupadas en un factor. Por el contrario, casi siempre se entienden los factores como simples resúmenes abstractos de las variables que se agrupan. Por lo general, cuando dos variables tienen altas cargas en el mismo factor, no se supone que exista una causa común que explique las dos variables, o que una sea causa de la otra, sino que se interpreta como un indicador de que las dos son distintas medidas de la misma característica. En esto se basa lo que McDonald (1981) denomina «heurística empírica». Esta práctica es muy común, y se resume en la frase «la naturaleza de un factor es aquello que tienen en común las variables que tienen altas cargas en el mismo». Esta práctica tautológica tiene el grave riesgo de hacernos caer en lo que Cliff (1983) llamó la falacia del nominalismo. Una cosa no es conocida sólo porque le demos un nombre.

En definitiva, debemos afirmar que cuando hablamos de unidimensionalidad no estamos postulando la singularidad conceptual. Se trata de cosas distintas como hemos visto. Y por otra parte, la comprensión de la naturaleza de la «unidimensión», supone un esfuerzo de elaboración teórica y de análisis en profundidad de sus relaciones con otras variables, esfuerzo que excluye la utilización de la heurística empírica para evitar el riesgo de caer en nominalismos tautológicos.

Conclusiones

En consecuencia, ¿qué consideraciones deben tenerse en cuenta al construir pruebas de estructura compleja que han de ser sometidas a un proceso de equiparación, tanto vertical como horizontal?

En primer lugar, respecto de lo equiparable, diremos que dos pruebas son equiparables si, dado un conjunto fijo de sujetos, la ordenación que produciría cada una de ellas por separado es la misma. Según esto, no importa tanto que la estructura interna de las pruebas sea compleja, como que el nivel y tipo de complejidad sea la misma para las dos. Esto supone que debemos tratar de mantener las pruebas tan estables como sea posible a lo largo del tiempo (Para observar el movimiento, la referencia debe permanecer inmóvil). Es especialmente importante mantener invariante la estructura de las pruebas, de modo que pueda asegurarse la constancia de la estructura factorial subyacente

En los casos en los que sea posible, debe tratarse de definir perfectamente, con la máxima precisión, las magnitudes relevantes que serán objeto de la medida. Cuando esas magnitudes estén claramente identificadas se deben producir estimaciones separadas para cada uno de los factores independientes.

Otra consecuencia importante es que debe evitarse producir una puntuación global, dado que en la práctica carece de significado operativo. Y naturalmente cada nueva materia a evaluar debe producir al menos una nueva escala independiente.

Por último, es muy probable que cambios en el currículo escolar o en la organización del sistema requieran introducir a su vez cambios en la evaluación. Cuando haya de introducirse cambios en las pruebas, tanto en su contenido como en su estructura, éstos habrán de llevarse a cabo progresivamente y previa comprobación experimental de la estabilidad de los resultados (prueba nueva en paralelo con prueba antigua). El objetivo debe ser siempre lograr la máxima comparativa de lo resultados año tras año.

Los modelos «psicométricos» actuales son tremendamente poderosos, pero es necesario conocer bien cuáles son las exigencias que introducen para que siempre podamos estar seguros de que cuando medimos la evolución y el cambio, las variaciones que observamos se deben a los datos y no a los instrumentos que utilizamos para medirlas.

Referencias bibliográficas

American Educational Research Association; American Psychological Association; National Council on Measurement in Education (AERA, APA, NCME) (1985): *Standards for Educational and Psychological Testing* . Washington DC.

Cliff, N. (1983): «Some Cautions Concerning The Application Of Causal Modelling Methods», en *Multivariate Behavioural Research*, Vol. 18, pp. 115-128.

Embretson, S. E.; Reise, S. P. (2000): *Item Response Theory for Psychologists.* Mahwah, New Jersey, Lawrence Erlbaum Associates.

Gaviria, J. L. (1988): *El supuesto de unidimensionalidad en la Teoría del Rasgo Latente.* Tesis Doctoral. Universidad Complutense de Madrid.

Hambleton, R. K.; Rovinelli, R. (1986): «Assessing the dimensionality of a set of items», en *Applied Psycological Measurement,* 10, pp. 287-302.

Hattie, J. (1984): «An empirical study of various indices for determining unidimensinality», en Multivariate Behavioral Research, 19, pp. 49-78.

— (1985): «Methodological Review: Assessing Unidimensionality of Tests and Items», en *Applied Psychological Measurement,* 9, pp. 139-164.

Kolen, M. J.; Brennan, R. L. (1995): *Test Equating. Methods and Practices.* New York. Springer Verlag.

Lord, F. M. (1980): *Applications of Items Response Theory to Practical Testing Problems.* Hillsdale, New Jersey, Lawrence Erlbaum Associates.

Lord, F. M.; Novick, M. R. (1968): *Statistical Theories of Mental Test Scores.* Reading, Mass. Addison-Wesley.

Lumsden, J. (1976): «Test Theory», en *Rosenzweig,* M. R.; *Porter,* L. W. (eds.): *VOL. Annual Review of Psychology.* Palo Alto, CA, Annual Reviews INC.

Martínez, F. (2000): «El sistema nacional de evaluación educativa de México (SNEE)», en Revista de Educación, 321, pp. 35-40.

Martínez, F.; Schmelkes, S. (2000): *Aseguramiento de la calidad de las pruebas de estándares nacionales para la Educación Primaria, de la secretaría de educación pública.* Ponencia presentada en el V Congreso de la Investigación Educativa. México.

Martínez- Rizo, F. (2004): «Compatibilidad de los resultados de las evaluaciones», en *Memorias de las Jornadas de Evaluación Educativa.*

Mcdonald, R. P. (1981): «The Dimensionality of Tests And Items», en British Journal Of Mathematical And Statistical Psychology, vol. 34, pp. 100-117

Mislevy, R. (1992): *Linking Educational Assessments: Concepts, issues, methods, and prospects.* Princeton, NJ, ETS Policy Information Center.

Mislevy, R.; Jhonson , E.; Muraki, E. (1992): «Scaling procedures in NAEP», en *Journal of Educational Statistics,* 17, pp. 131-154.

Muraki, E.; Engelhard, G. (1985): «Full information factor análisis: applications of EAP scores», en *Applied Psychological Measurement,* 9, pp. 417-430.

Muñiz, J. (1990): *Teoría clásica de los test.* Madrid, Pirámide.

Rasch, G. (1960): *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen, Danish Institute for Educational Research.

Secretaría de Educación Pública (2002): *La Experiencia de la Dirección General de Evaluación en la Educación Básica y Normal. 30 años de medición del logro educativo.* México, Dirección General de Evaluación del Desempeño.

Stout, W. (1996): *Dimensionality-Based DIF/DBF Package: Sibtest, Poly-Sibtest, Crossing Sibtest Difsim, Difcomp. IRT-Based Educational and Psychological Measurement Software.* The William Stout Institute for Measurement.

Páginas web

www.sep.gob.mx/work/appsite/dge/index.htm
www.senado.gob.mx/.../educacion/content/documentos/cuadernos_trabajo/seminario/tema 2/velazquez/index.pdf)