Validez y fiabilidad de un cuestionario sobre medidas de tendencia central para estudiantes de secundaria y bachillerato

Silvia Azucena **Mayén** Galicia
Departamento de Matemática Educativa, Centro de Investigación y de Estudios Avanzados
México
mayazuc@gmail.com
Carmen **Díaz** Batanero
Departamento de Psicología, Universidad de Huelva
España
carmen.diaz@dpsi.uhu.es

Resumen

Presentamos el estudio de validez y fiabilidad, de un cuestionario de medidas de tendencia central construido por Cobo (2003), orientado a evaluar la comprensión que sobre estos conceptos tienen estudiantes mexicanos de secundaria y bachillerato. Se aplicaron distintos métodos estadísticos: análisis factorial, clúster e implicativo para determinar las dimensiones del cuestionario y analizar la explicación teórica de estas dimensiones. Se complementa con métodos bayesianos con la finalidad de proporcionar información útil en nuevas investigaciones o en la evaluación de procesos de instrucción y otros usos en investigación didáctica. La finalidad es obtener información empírica sobre las características de este instrumento y demostrar que puede ser generalizable a estudiantes de otros contextos y niveles de estudios.

Palabras clave: Medidas de tendencia central, validez, fiabilidad, estudiantes de secundaria, bachillerato.

Introducción

A partir de un estudio de evaluación con una muestra de 518 estudiantes mexicanos: 356 de Bachillerato y 162 de Secundaria, y mediante un cuestionario que evalúa la comprensión de las medidas de tendencia central, se lleva a cabo en este trabajo, la validación de este instrumento, construido por Cobo (2003), recogiendo tres tipos de evidencia de la validez: validez de contenido (justificada mediante análisis teórico de los ítems); validez discriminante (mediante análisis de diferencia de ejecución en los ítems en los grupos); y validez de constructo (analizando la estructura de las respuestas mediante análisis cluster e implicativo). La aproximación a la fiabilidad se lleva a cabo mediante el coeficiente Alfa (Martínez Arias, 1995), coeficiente Theta (Barbero, 2003) y Teoría de la Generalizabilidad (Feldt y Brennan, 1991). Se realiza un estudio global de la dificultad de los ítems del cuestionario en nuestra muestra y una comparación ítem a ítem de los resultados en los dos grupos de estudiantes.

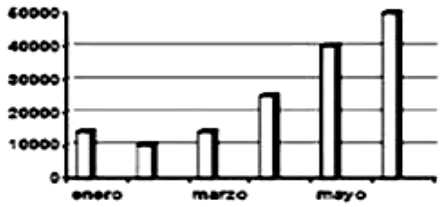
El cuestionario

Un *cuestionario* es un instrumento de medición que por medio de las preguntas planteadas obtiene una estimación de conocimientos y capacidades de los sujetos a quienes se les aplica, que no son accesibles por simple observación o encuesta (Dane, 1990; Barbero, 2003). Al tratar de evaluar la comprensión, se tuvo en cuenta que es un constructo inobservable (León y Montero, 2002), por lo que sus características deben ser inferidas de las respuestas de los alumnos. La comprensión de los estudiantes sobre un cierto objeto matemático (en este caso las medidas de tendencia central) es inobservable. Pero las prácticas que realizan al resolver problemas, y en particular los problemas presentados como ítems en un cuestionario, sí que son observables, siempre que la recopilación de datos sea completa y fiable (Godino, 1996).

Este cuestionario tiene, por tanto, como principal objetivo, recoger datos sobre las prácticas matemáticas que realizan los estudiantes al resolver problemas relacionados con las medidas de tendencia central. De las respuestas escritas trataremos de inferir el uso (correcto o incorrecto) que los estudiantes de la muestra hacen de los diversos objetos matemáticos (definiciones, propiedades, argumentos, etc.). Las interpretaciones realizadas a partir de las respuestas harán referencia a lo que los sujetos hacen o son capaces de hacer, y a sus conocimientos y errores sobre el mismo (*test referido a criterio*). Se trata entonces de un cuestionario de *potencia*, ya que el tiempo, aunque controlado, no determinaría el resultado, sino que las diferencias en la puntuación serían debidas a la calidad de su ejecución y conocimiento (Sax, 1989; Martínez Arias, 1995). El cuestionario fue construido por Cobo (2003), después de un análisis sistemático del contenido de las medidas de posición central en una amplia muestra de libros de texto españoles de secundaria, así como del currículo español. Se compone de dieciséis ítems abiertos que presentan enunciados de situaciones comprensibles y familiares para los estudiantes. Para nuestro caso, también hemos verificado que los contenidos se incluyan en los programas de estudios mexicanos de enseñanza Secundaria y Media Superior (DEMS, 1997; SEP, 2006), que a continuación resumimos: campos de problemas que se resuelven mediante promedios; diferentes definiciones de media, mediana y moda; comprensión de propiedades numéricas, algebraicas y estadísticas; reconocimiento del lenguaje matemático verbal, numérico y gráfico y uso apropiado de términos y lenguaje; cálculo y procedimientos de resolución de problemas; y argumentos de los alumnos para apoyar sus respuestas. A continuación presentamos los ítems:

1. Un periódico dice que el número medio de hijos por familia en México es 2.2 hijos por familia. Explica qué significa para ti esta frase. Se han elegido 10 familias mexicanas y el número medio de hijos entre las 10 familias es de 2.2 hijos por familia. Los García tienen 4 hijos y los Pérez tienen 1 hijo, ¿cuántos hijos podrán tener las otras 8 familias para que la media de hijos en las 10 familias sea 2.2? Explica tu respuesta.
2. María y Pedro dedican una media de 8 horas cada fin de semana a hacer deporte. Otros 8 estudiantes dedican cada fin de semana una media de 4 horas a hacer deporte. a) ¿Cuál es el número medio de horas que hacen deporte cada fin de semana los 10 estudiantes? María y Pedro dedican además 1 hora cada fin de semana a escuchar música y los otros 8 estudiantes, 3 horas. b) ¿Cuál es el número medio de horas que escuchan música los diez estudiantes?, c) ¿Cuál sería el número medio de horas que estos 10 estudiantes dedican cada fin de semana, entre las dos actividades: hacer deporte y escuchar música?
3. Cuatro amigos se reúnen para preparar una cena. Cada uno de ellos trajo harina para hacer la masa de las pizzas. Como querían hacer cuatro pizzas del mismo tamaño, los que habían traído más harina regalaron a los que llevaban menos. ¿La cantidad de harina regalada por los que habían traído mucha fue ***mayor, menor o igual*** a la recibida por los que habían traído poca? ¿Por qué piensas eso?

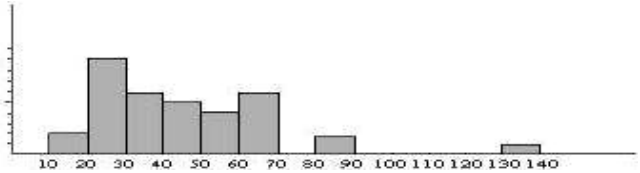
4. Tenemos ***seis números*** y el más grande es el 5. Sumamos estos números y dividimos la suma entre ***seis***. El resultado es 4. ¿Te parece posible? ¿Por qué?
5. El peso en kilos de 9 niños es 15, 25, 17, 19, 16, 26, 18, 19, 24. ¿Cuál es el peso del niño **mediano**? ¿Cuál es la **mediana** si incluimos el peso de otro niño que pesa 43 Kg? En este caso, ¿Sería la **media aritmética** un buen representante de los 10 datos?, ¿Por qué?
6. Un profesor califica a sus alumnos del siguiente modo: I=Insuficiente, A=Aprobado, N=Notable, S=Sobresaliente. A continuación tenemos las notas que ha puesto a dos grupos de alumnos:
   Grupo 1: I A A N N S S I I I A A A N S S I A A S S S S; Grupo 2: S S I I A N A N I I S N A S I N N
   a) ¿Qué grupo ha obtenido mejores notas? b) ¿Cuál sería el promedio (medida de centralización) más apropiado para representar estos datos? Explica tu respuesta.
7. Lucía, Juan y Pablo van a una fiesta. Cada uno lleva cierto número de caramelos. Entre todos llevan una media de 11 caramelos por persona. ¿Cuántos caramelos ha llevado cada uno? Lucía__ Juan__ Pablo ___
   b) ¿Es la única posibilidad? Si__ No__ Explica cómo has obtenido tus resultados. Un cuarto chico llega a la fiesta y no lleva ningún caramelo. ¿Cuál es ahora la media de caramelos por chico? Explica.
8. Nueve estudiantes han pesado un objeto en la clase de ciencias, usando la misma escala. Los pesos registrados por cada estudiante (en gramos) se muestran a continuación: 6.2, 6.3, 6.0, 15.2, 6.2, 6.1, 6.5, 6.2, 6.1, 6.2. Los estudiantes quieren determinar con la mayor precisión posible el peso real del objeto. ¿Qué harías para calcularlo?
9. Observa el siguiente diagrama de barras que muestra las ventas de bocadillos de la empresa *Bocatta* durante los últimos 6 meses del año pasado:



   a) Da un valor aproximado del número medio de bocadillos que se vende al mes.
   b) Da un valor aproximado de la mediana del número de bocadillos que se vendieron por mes.
10. El siguiente conjunto de datos refleja las edades en que contrajeron matrimonio una muestra de 100 mujeres.

Edad	Frecuencia
15-19	4
20-24	38
25-29	28
30-34	20
35-39	8
40-44	1
45-49	1

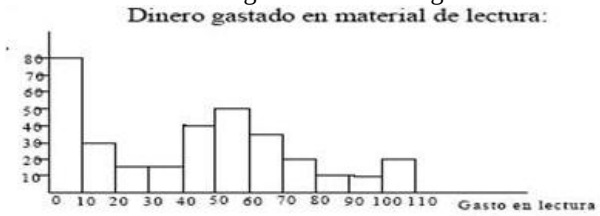
    ¿Cuál es la media, mediana y moda de la edad de estas mujeres? Realiza los cálculos necesarios.
11. Pedro piensa que en el siguiente gráfico, la mediana te dice que una mayoría de estudiantes gasta alrededor de 47 euros cada semana en comida.



Gasto semanal en comida de un grupo de estudiantes
Media = 45.3 euros

    Si _____ No_____ ¿Por qué?
12. Juana piensa que, en el gráfico anterior, el alto valor de 133 euros debería quitarse del conjunto de datos antes de calcular media, mediana y moda. Si____ No ____¿Por qué?

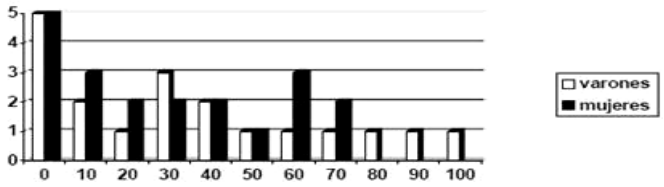
13. Un grupo de estudiantes hizo este otro histograma del dinero gastado cada semana en material de lectura.



Media = 29 Euros. Mediana = 25 Euros

Francisco dijo que es difícil describir el dinero típico gastado en material de lectura en el gráfico anterior, porque la mayor parte de los estudiantes no gastaron nada o muy poco y un segundo grupo gastó entre 50 y 60 euros cada semana. Él piensa que en este caso la media es un indicador pobre del promedio y elige usar la mediana en su lugar para representar el gasto semanal "promedio" en material de lectura.
Si_____ No_____ ¿Por qué?

14. Manuel dice que esta distribución tiene dos modas y sugiere que puedes obtener más información de estos datos si los divides en dos subgrupos y calculas el promedio en cada grupo por separado.
Si_____ No _____ ¿Por qué?
15. Lola sugiere hacer el estudio en pesos mexicanos (1€ = $14.5). ¿Cuál sería en este caso el valor de la media?
16. Antonio quiere investigar las diferencias en los hábitos de gasto en varones y mujeres. Compara las cantidades gastadas en material de lectura en varones y mujeres, construyendo los gráficos siguientes:



Antonio cree que sus gráficos muestran que los varones y mujeres tienden a gastar diferentes cantidades de dinero en material de lectura. Si_____ No _____ ¿Por qué?

Validez de contenido del cuestionario

La *validez de contenido del cuestionario* es el grado en que el instrumento de evaluación refleja el dominio que nos interesa en forma satisfactoria (Carmines y Zeller, 1979). Se trata de ver la adecuación de los ítems de un test como muestra de un universo más amplio de ítems representativos del contenido (Martínez Arias, 1995). La validación del contenido se ha hecho mediante el examen sistemático del contenido del test para probar su representatividad y relevancia. Con ello comprobamos que los ítems del test son relevantes para el uso que se dará a las puntuaciones y representativos del contenido que se quiere evaluar, resaltando sus características esenciales.

Estudio global de resultados

En este estudio presentamos los resultados obtenidos de las respuestas de los estudiantes al cuestionario administrado, para la totalidad de la muestra y comparando los grupos que la componen, es decir, entre estudiantes de secundaria y bachillerato. Se organiza en los siguientes apartados: a) análisis comparado de la dificultad de los ítems; b) cálculo de los intervalos de confianza y credibilidad para estos índices (estimaciones clásicas y estimaciones bayesianas); c) análisis de la puntuación total; y d) estudio de fiabilidad y generalizabilidad del cuestionario.

Dificultad comparada de ítems. Para realizar el análisis de *dificultad de los ítems* del cuestionario tomamos las respuestas obtenidas de los estudiantes, y observamos si el alumno es capaz o no de

dar las respuestas correctas esperadas, y categorizamos las respuestas como correctas o incorrectas. En la Tabla 1 se presentan para el total de la muestra los *índices de dificultad* de cada ítem, entendiéndolo, según Muñiz (1994), como la proporción de sujetos que lo aciertan entre todos los que trataron de resolverlo. Cuanto mayor es este valor, significa que el ítem es más fácil y ha sido respondido correctamente por una mayor proporción de estudiantes. El índice fluctúa entre 0.24 en el ítem 2.3 (*media de una suma de variables*), 0.26 en el ítem 10.1 (*media, mediana y moda de un conjunto de datos agrupados en intervalos presentados en una tabla de frecuencias*), y 0.85 en el ítem 8 (*estimación de una cantidad desconocida a partir de diversas mediciones en presencia de errores*). La mayor parte de los ítems tiene dificultad moderada. En concreto, 24 de los 27 subítems tienen un índice de dificultad comprendido entre 0.3 y 0.7. Con ello conseguiremos más discriminación entre los estudiantes, y en definitiva mejores resultados en la evaluación. Las desviaciones típicas, se agrupan alrededor de 0.50, lo que representa una variabilidad elevada para el caso de ítems dicotómicos. Es decir, los ítems permitirán mostrar, en caso de que existan, las diferencias individuales en la dificultad encontrada por los alumnos con respecto a la media.

Tabla 1. Índice de dificultad y desviación típica (n=518)

Ítem	Índice de dificultad	Desviación típica
i1_1	0.71	0.453
i1_2	0.65	0.478
i2_1	0.30	0.458
i2_2	0.37	0.484
I2_3	0.24	0.427
i3	0.65	0.476
i4	0.76	0.426
i5_1	0.68	0.466
i5_2	0.60	0.490
i5_3	0.31	0.462
i6_1	0.37	0.482
i6_2	0.36	0.480
i7_1	0.78	0.416
i7_2	0.76	0.428
i7_3	0.68	0.466
i8	0.85	0.356
i9_1	0.63	0.484
i9_2	0.32	0.466
i10_1	0.26	0.439
i10_2	0.30	0.458
i10_3	0.59	0.491
i11	0.34	0.475
i12	0.27	0.443
i13	0.37	0.483
i14	0.34	0.475
i15	0.57	0.496
i16	0.35	0.478

Estimaciones bayesianas. Una de las razones sobre la conveniencia de que los métodos clásicos de inferencia sean sustituidos o complementados con métodos bayesianos, es la interpretación más intuitiva de los resultados proporcionados y, sobre todo, la posibilidad

de tener en cuenta la información previa que se posea sobre la población en estudio. En nuestro caso, utilizamos la información previa proporcionada por Cobo (2003), que ofrece un análisis detallado de la dificultad de estos ítems en dos muestras de alumnos de Secundaria en España, siendo muy semejantes en edad y cursos estudiados a nuestro grupo de estudiantes de Secundaria, además de los contenidos muy similares en ambos países, por lo que cabe esperar que la dificultad de cada ítem sea próxima a la que se obtuvo en dicho estudio y que la dificultad relativa de los ítems se conserve en los dos grupos de estudiantes. Aquí presentamos la estimación clásica y bayesiana de los índices de dificultad de los ítems que conforman el cuestionario. De esta forma mejoraremos nuestras estimaciones, además de ofrecer una interpretación más natural de los intervalos en torno a estas estimaciones.

La Tabla 2 presenta estimaciones clásica y bayesiana de los índices de dificultad de los ítems del cuestionario. Incluye intervalos de confianza (estimación clásica) e intervalos de credibilidad (estimación bayesiana) para los índices de dificultad de cada uno de los ítems.

Tabla 2. Índice de dificultad, Intervalos de Confianza y Credibilidad del 95%

	Estimación Clásica	Intervalo de confianza		Intervalo de credibilidad No informativo	
Ítem	Índice dificultad	L. inferior	L. superior	L. inferior	L. superior
I1_1	0.71	0.671	0.749	0.670	0.748
I1_2	0.65	0.610	0.692	0.608	0.690
I2_1	0.30	0.260	0.339	0.261	0.340
I2_2	0.37	0.329	0.412	0.330	0.413
I2_3	0.24	0.203	0.276	0.204	0.278
I3	0.65	0.610	0.692	0.608	0.690
I4	0.76	0.724	0.797	0.722	0.795
I5_1	0.68	0.639	0.720	0.639	0.719
I5_2	0.60	0.558	0.643	0.558	0.642
I5_3	0.31	0.271	0.351	0.272	0.351
I6_1	0.37	0.329	0.412	0.330	0.413
I6_2	0.36	0.318	0.400	0.319	0.401
I7_1	0.78	0.744	0.816	0.743	0.814
I7_2	0.76	0.724	0.797	0.722	0.795
I7_3	0.68	0.639	0.720	0.639	0.719
I8	0.85	0.819	0.880	0.817	0.879
I9_1	0.63	0.588	0.671	0.588	0.671
I9_2	0.32	0.280	0.361	0.281	0.361
I10_1	0.26	0.223	0.298	0.224	0.299
I10_2	0.30	0.260	0.339	0.261	0.340
I10_3	0.59	0.548	0.633	0.548	0.632
I11	0.34	0.299	0.381	0.300	0.381
I12	0.27	0.671	0.749	0.670	0.748
I13	0.37	0.610	0.692	0.608	0.690
I14	0.34	0.260	0.339	0.261	0.340
I15	0.57	0.329	0.412	0.330	0.413
I16	0.35	0.203	0.276	0.204	0.278

El primero ha sido calculado con la fórmula ordinaria de intervalos de confianza de una proporción con interpretación frecuencial, es decir, en cada 100 muestras tomadas de la misma población, 95% de ellas contendrían la proporción verdadera, aunque no podemos saber si se contiene o no en nuestra muestra. El intervalo de credibilidad, por el contrario, indica el intervalo de valores en que esperamos que la proporción verdadera esté incluida, es decir, nos da una probabilidad epistémica, que se refiere a la muestra particular.

En la Tabla 3 presentamos las estimaciones bayesianas, en este caso con una distribución inicial informativa (usando la información del estudio de Cobo, 2003). Observamos que ahora los intervalos de credibilidad son más precisos y el valor estimado se corrige con la información previa, aunque, como nuestra muestra es mayor, en caso de diferencia entre las dos estimaciones, el valor final se aproxima más al de nuestro estudio.

Tabla 3. *Estimación bayesiana de índices de dificultad con distribución informativa*

			Estimación bayesiana del índice de dificultad		
	Índice observado n=518	Proporción en Cobo (2003), 4ºESO, n=144	Valor medio	L. inferior	L. superior
I1_1	0.71	0.69	0.706	0.670	0.740
I1_2	0.65	0.37	0.589	0.551	0.626
I2_1	0.30	0.34	0.308	0.274	0.344
I2_2	0.37	0.38	0.373	0.336	0.410
I2_3	0.24	0.33	0.260	0.227	0.294
I3	0.65	0.49	0.616	0.578	0.652
I4	0.76	0.66	0.738	0.704	0.771
I5_1	0.68	0.38	0.615	0.577	0.651
I5_2	0.60	0.32	0.539	0.501	0.577
I5_3	0.31	0.33	0.315	0.280	0.351
I6_1	0.37	0.13	0.318	0.284	0.354
I6_2	0.36	0.04	0.290	0.256	0.325
I7_1	0.78	0.67	0.756	0.722	0.788
I7_2	0.76	0.68	0.743	0.709	0.775
I7_3	0.68	0.61	0.665	0.628	0.700
I8	0.85	0.67	0.811	0.780	0.840
I9_1	0.63	0.67	0.638	0.601	0.674
I9_2	0.32	0.26	0.307	0.272	0.342
I10_1	0.26	-	0.261	0.224	0.299
I10_2	0.30	-	0.300	0.261	0.340
I10_3	0.59	-	0.590	0.548	0.632
I11	0.34	0.26	0.322	0.287	0.358
I12	0.27	0.15	0.245	0.213	0.278
I13	0.37	0.34	0.364	0.328	0.401
I14	0.34	0.27	0.325	0.290	0.361
I15	0.57	0.45	0.544	0.506	0.582
I16	0.35	0.49	0.381	0.344	0.418

Estudio de fiabilidad y generalizabilidad. El análisis de la *fiabilidad* del cuestionario en nuestra muestra, se entiende como la extensión por la cual un experimento, test u otro procedimiento de medida produce los mismos resultados en ensayos repetidos. La medida siempre produce un cierto error aleatorio, pero dos medidas del mismo fenómeno sobre un mismo individuo suelen ser consistentes. Sin embargo, la fiabilidad varía al cambiar la población de estudio. Siguiendo a Thorndike (1989), evaluamos conceptos abstractos, en decir, comprensión de los alumnos sobre medidas de posición central con sus respuestas a los ítems del cuestionario, que son indicadores empíricos. Para permitir este proceso de medida, un indicador debe ser fiable. La *fiabilidad* es la tendencia a la consistencia o precisión del instrumento en la población medida (Bisquerra, 1989). Resaltamos que para el cálculo de fiabilidad y generalizabilidad usamos todos los datos de la muestra descrita e incluimos otra muestra de 125 estudiantes mexicanos, que formaron parte de un estudio piloto (Mayén, 2006), con un total de 643 estudiantes. El objetivo es tener una estimación más precisa de la fiabilidad al contar con una muestra de mayor tamaño. Consideramos el método de *consistencia interna* para estimar la fiabilidad de una escala, que está basado sólo en la aplicación del cuestionario (Díaz, Batanero y Cobo, 2003), el coeficiente que lo mide es el Alfa de Cronbach (Carmines y Zeller, 1979), del cual se obtuvo un valor Alfa = 0.662, que se considera como un valor adecuado aunque no excesivamente elevado debido a que el cuestionario evalúa un constructo que no es unidimensional, es decir, que incluye una gama amplia de conceptos. También se calculó un coeficiente de fiabilidad basado en el análisis factorial. Puesto que asumimos que el cuestionario evalúa un constructo multidimensional, la fiabilidad se calcula más exactamente con el coeficiente Theta de Carmines, que representa la contribución del primer factor del análisis factorial al total de la fiabilidad del cuestionario. A partir de los resultados del análisis factorial se calculó el coeficiente Theta de Carmines[1]:

$$\theta = \frac{n}{n-1}\left(1-\frac{1}{\lambda_1}\right)=0.726$$

donde *n* es el número de ítems y λ el primer autovalor en el análisis factorial. Coincide con el coeficiente α calculado con las puntuaciones factoriales derivadas del primer factor común y sintetiza la información aportada por el primer factor (Morales, 1988). El coeficiente alfa presenta un valor bastante alto.

Coeficientes de generalizabilidad. La teoría de la generalizabilidad extiende la teoría clásica de la medición, (Feldt y Brennan, 1991), y permite por medio del análisis de varianza, observar diferentes fuentes de error en un proceso de medida. El *coeficiente de generalizabilidad* se define con el cociente (1), es decir, como el cociente entre la varianza verdadera en las puntuaciones de la prueba y la varianza observada, que es la suma de la varianza verdadera más la varianza debida al error aleatorio. En este trabajo hemos diferenciado dos fuentes para el error aleatorio y calculado la generalizabilidad de los mismos sujetos (inter-personas) y la generalizabilidad de los ítems (inter-elementos).

$$G = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2} \quad (1)$$

[1] Carmines y Zéller (1979) definen este coeficiente para cuestionarios no unidimensionales, como es el nuestro.

Sustituyendo ahora estos componentes de varianza en la fórmula (1) y teniendo en cuenta los tamaños de muestra (27 ítems y 643 alumnos), según si consideramos como fuente de variación los problemas o los alumnos, obtenemos las siguientes estimaciones:

Generalizabilidad respecto a los ítems: $$G_i = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2 / 27} = 0.6616$$

Obtenemos un valor próximo al del coeficiente Alfa (0.662), debido a que el coeficiente de generalizabilidad respecto a los ítems coincide con él, ya que se considera el número de alumnos fijo y la única fuente de variación se debe a la variabilidad entre ítems.

Generalizabilidad respecto a los alumnos: $$G_s = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_e^2 / 643} = 0.9619$$

El valor para la generalizabilidad es muy alto respecto a otros alumnos de la misma prueba, es decir, se indica una muy alta posibilidad de generalizar nuestros resultados a otros alumnos conservando el mismo cuestionario, por supuesto, bajo la hipótesis de que se conserven las características sociológicas y educativas.

Estructura de las respuestas. Para estudiar las interrelaciones entre objetivos de aprendizaje hemos llevado a cabo varios análisis multivariantes de las respuestas a los ítems de la prueba. Los resultados de estos análisis se presentan a continuación.

Análisis cluster. Con el software CHIC, Classification Hierarchical, Implicative et Cohesive (Couturier y Gras, 2005), se realiza un estudio de aglomeración jerárquica, tomando como medida de similaridad entre ítems el índice de Lerman y suponiendo una distribución binomial para cada variable. Siendo *a* y *b* dos variables aleatorias dicotómicas en una población, *E* y *A* y *B* los subconjuntos donde se verifican *a* y *b*, el índice de similaridad viene dado por la expresión siguiente (Lerman, 1981):

$$\partial(a,b) = \frac{card(A,B) - \frac{card(A)card(A,B)}{n}}{\sqrt{\frac{card(A)card(A,B)}{n}}}$$

Las variables *a* y *b* tendrán mayor similaridad cuando el número de elementos comunes sea mayor en relación a la frecuencia esperada en caso de independencia y tiene en cuenta el tamaño muestra. En nuestro caso *A* representa el conjunto de estudiantes que contesta el ítem *a* y *B* el que responde el ítem *b*, *(A,B)* el conjunto de los que responden correctamente a los dos ítems. El programa CHIC proporciona también una prueba de significación de las clases obtenidas, que se calcula teniendo en cuenta no sólo la intensidad de la similaridad, sino el número total de sujetos que da una respuesta correcta al conjunto de ítems incluidos en el grupo.

En la Figura 1 presentamos el dendograma que muestra los grupos formados de ítems donde los mismos alumnos dan contestaciones similares (bien correctas, bien incorrectas). Es decir, se trata de conocimientos relacionados entre sí y separados de los otros grupos de ítems. Observamos una estructura muy compleja con numerosos grupos, lo que indica componentes diferenciados en el significado de las medidas de tendencia central.

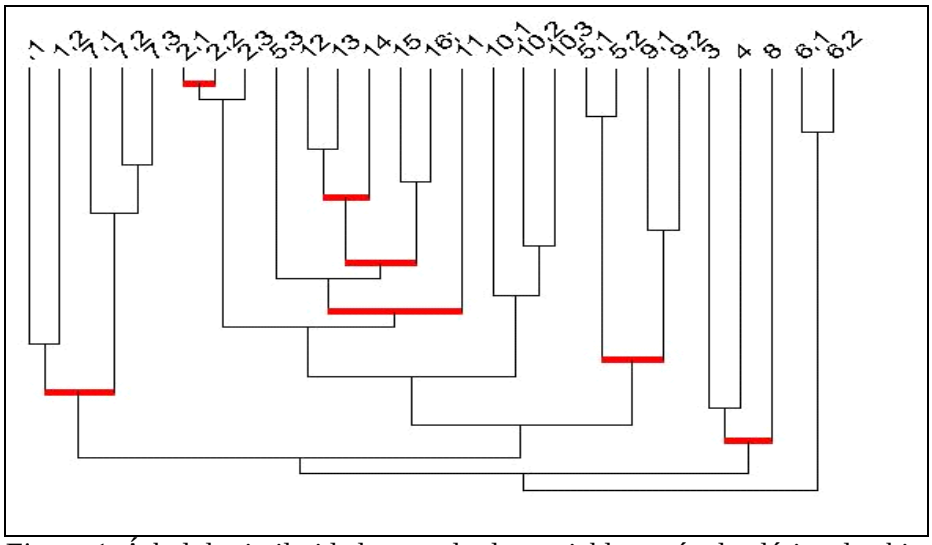


Figura 1. Árbol de similaridad con todas las variables, método clásico, ley binomial.

Grafo implicativo. El análisis cluster presentado considera una medida de asociación simétrica en las variables, es decir, se supone que si un estudiante responde a un ítem también responderá a otro asociado con él, pero no tiene en cuenta la dificultad relativa de cada ítem. Una situación más plausible es pensar que, aunque dos ítems estén relacionados, si uno es más difícil, la respuesta a este ítem facilita que también se acierte en el segundo. El análisis implicativo entre ítems, proporciona un estudio de la implicación (no simétrica) entre el conjunto de ítems, es decir, se trata de ver si la respuesta correcta al ítem a implica la respuesta correcta al *b* (donde la respuesta correcta a *b* puede o puede que no implique la respuesta a *a*), (Figura 2).

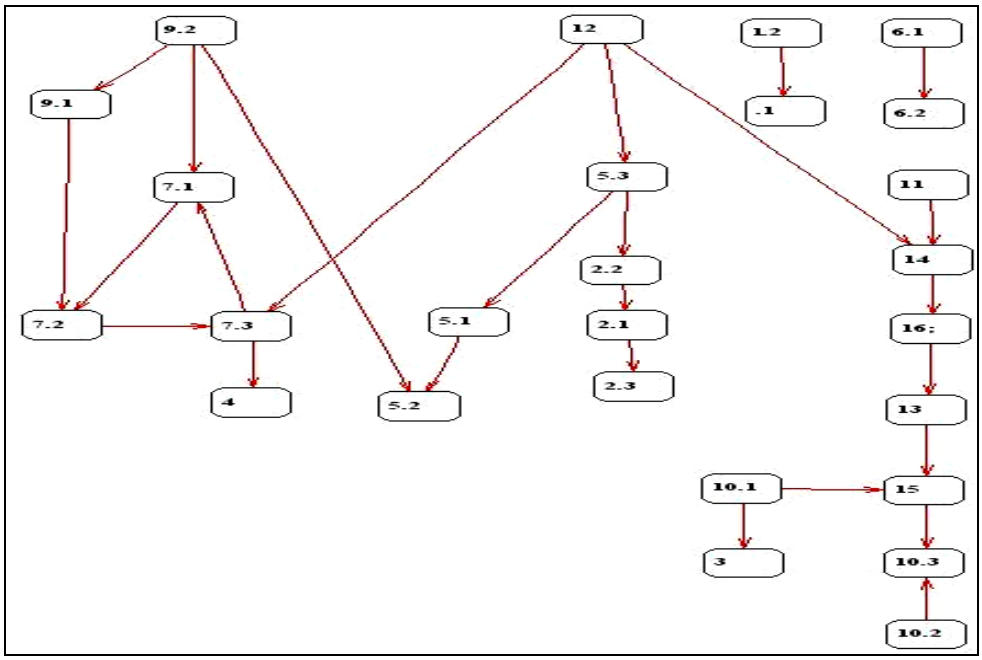


Figura 2. Grafo implicativo.

La implicación entre un ítem y otro se interpreta en el sentido de que si un estudiante es capaz de resolver correctamente un ítem, entonces mejora su probabilidad de resolver correctamente otro implicado por aquél. En este sentido, el árbol implicativo nos proporciona una pauta de posible orden de introducción de los conceptos y procedimientos evaluados por los diferentes ítems. Hacemos énfasis en que la relación de implicación es asimétrica, indicándose el sentido de la implicación por la dirección de la flecha en el grafo. Por tanto, si estudiamos las relaciones significativas al 99%, que aparecen en rojo, la interpretación es que los alumnos resuelven correctamente el ítem. Como ejemplo, proporcionamos una interpretación de un grupo de ítems implicados:

- El alumno que es capaz de determinar la mediana a partir de un gráfico (ítem 9.1) también será capaz de hallar la media porque en los dos casos tiene que interpretar un gráfico y el concepto de mediana es más complejo que el de media. Por otro lado, el estudiante que calcula la mediana a partir del gráfico también realiza mejor el problema de determinar una distribución (ítem 7.1) y dar un segundo ejemplo (ítem 7.2), posiblemente porque la interpretación correcta de un gráfico implica la comprensión de la idea de distribución representada en el gráfico.

En definitiva, el grafo implicativo apoya nuestra hipótesis de que la comprensión de medidas de tendencia central por los estudiantes no puede concebirse como un constructo unitario, lo que explica que el análisis factorial haya resultado con tantos factores y la fiabilidad (coeficiente Alfa) con un valor moderado. Por el contrario, el grafo implicativo muestra una jerarquía de conocimientos entrelazados que los alumnos deben conseguir progresivamente y el profesor debe considerar en la enseñanza del tema.

Análisis implicativo jerárquico. El grafo implicativo, aunque muestra la estructura de interrelaciones, es algo complejo, por lo que sería interesante tratar de dividir el conjunto de ítems en unos pocos grupos interrelacionados entre sí mediante el *índice de implicación.* Una vez estudiadas las implicaciones aisladas de unos ítems sobre otros, hemos llevado a cabo un estudio de clasificación implicativa. Se trata de un algoritmo que utiliza las intensidades de implicaciones entre conjuntos de variables como índice no simétrico para estudiar la cohesión interna de algunos subconjuntos de variables. La cohesión de una clase tiene en cuenta la cantidad de información proporcionada por un conjunto de variables, el índice se puede interpretar como cantidad de información que una variable proporciona sobre otra. El programa CHIC calcula el nivel de significación de los diferentes nodos en una jerarquía implicativa, así como las contribuciones de los sujetos. El algoritmo forma las clases teniendo en cuenta: a) la cohesión máxima dentro de cada clase, y b) el mayor grado de implicación entre una clase y otra que es implicada por ella. Completamos el estudio con la determinación de una jerarquía implicativa en el conjunto de variables y mostramos en la Figura 3 el árbol de cohesión implicativa, donde se observan cinco grandes grupos de ítems, explicamos sólo el primero de ellos:

1. *Cálculo avanzado de la media y comprensión procedimental.* Ítems 2.1, 2.2 y 2.3, que se implican todos ellos entre sí; 1.2, que implica a los ítems 7.1, 7.2, 7.3, que se implican entre sí e implican al ítem 4.

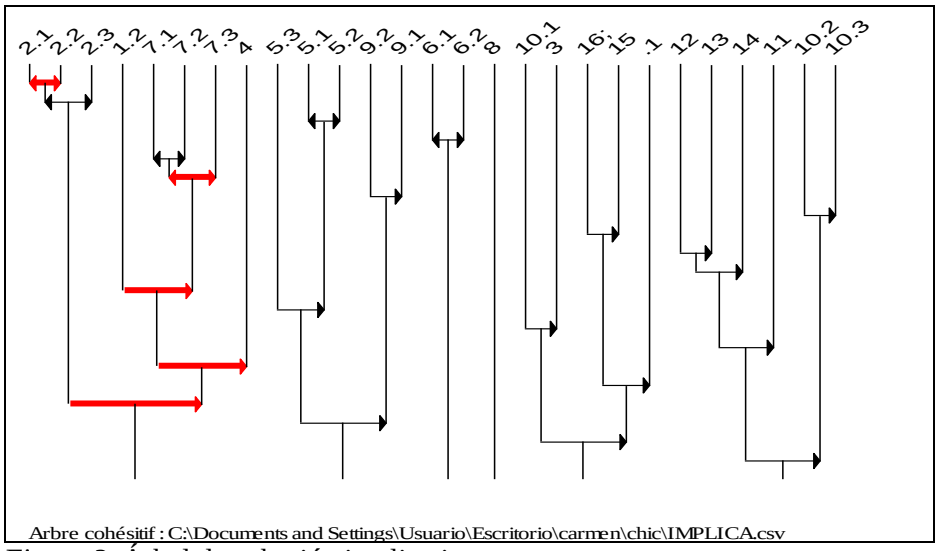


Figura 3. Árbol de cohesión implicativa.

Conclusiones

Nuestra principal aportación de este trabajo es el estudio de validación del cuestionario construido por Cobo (2003), que incluye validez de contenido, validez discriminante y de constructo. Hemos realizado también cálculos de índices de fiabilidad y generalizabilidad, que se completan con el coeficiente Theta basado en el análisis factorial, más adecuado para cuestionarios multidimensionales como es el nuestro. Proporcionamos estimaciones bayesianas de los índices de dificultad del cuestionario, muy apropiadas para situaciones con información previa. Al comparar la dificultad relativa de los diferentes ítems, encontramos semejanzas con investigaciones anteriores (Cobo, 2003; Mayén, 2006). Hay mayor facilidad para resolver problemas de media simple; utilizar la media como mejor estimador de una cantidad en presencia de errores de medida, y utilizar la moda. Hay también coincidencia en los ítems más difíciles, es decir, confundir media y mediana; evaluar la comprensión del efecto de valores atípicos sobre el cálculo de la media; no reconocer el efecto del cero sobre su cálculo; resolución de ítems de media y mediana con variables ordinales; y principalmente problemas de mediana, tanto en su cálculo como en su interpretación a partir de datos presentados en forma de gráficos y problemas de media ponderada. Añadimos la identificación de dificultades en el cálculo de la media, mediana y moda de un conjunto de datos agrupados en intervalos y presentados en una tabla de frecuencias absolutas.

Bibliografía

Barbero, M. (2003). *Psicometría II. Métodos de elaboración de escalas*. Madrid: UNED.

Batanero, C. y Díaz, M. C. (2005). Análisis del proceso de construcción de un cuestionario sobre probabilidad condicional. Reflexiones desde el marco de la TFS. En A. Contreras (Ed.), *Investigación en Didáctica de las Matemáticas* (pp. 13-36). Universidad de Jaén.

Bisquerra, R. (1989). *Métodos de investigación educativa.* Barcelona: CEAC.

Carmines, E. G. y Zeller, R. A. (1979). *Reliability and validity assesment.* Londres: Sage University Paper.

Cobo, B. (2003). *Significado de las medidas de posición central para los estudiantes de secundaria.* Tesis Doctoral. Universidad de Granada.

Couturier, R. y Gras, R. (2005). CHIC: Traitement de données avec l'analyse implicative. En G. Ritschard y C. Djeraba (Eds.), *Journées d'extraction et gestion des connaissances (EGC'2005)* (Vol. 2, pp. 679-684). Universidad de Lille.

Dane, F. C. (1990). *Research methods*. Thompson. Pacific Grow. CA.

DEMS (1997). *Programa de Estudios Probabilidad y Estadística.* Instituto Politécnico Nacional, Secretaría Académica, Dirección de Educación Media Superior, México.

Díaz, C. (2007). *Viabilidad de la inferencia bayesiana en el análisis de datos en psicología.* Tesis Doctoral. Universidad de Granada.

Díaz, C., Batanero, C. y Cobo, B. (2003). Fiabilidad y generalizabilidad. Aplicaciones en evaluación educativa. *Números, 54*, 3 – 21.

Dunn, O, J. y Clarck, V. A. (1987). *Applied statistics: Analysis of variance and regression.* Nueva York: John Wiley.

Feldt, L. S. y Brennan, R. L. (1991). Reliability. En R. L. Linn (Ed.), *Educational measurement.* (pp. 105-146). Nueva York: MacMillan.

Godino, J. D. (1996). Mathematical concepts, their meanings and understanding. En L. Puig y A. Gutiérrez (Eds.), *Proceedings of the 20th PME Conference* (v.2, 417-424). Universidad de Valencia.

Godino, J. D. (1999) *Análisis epistémico, semiótico y didáctico de procesos de instrucción matemática.* On line: www.ugr.es/~jgodino/.

León, O. G. y Montero, I. (2002). *Métodos de investigación en psicología y educación.* Madrid: McGraw-Hill.

Martínez Arias, R. (1995). *Psicometría: teoría de los tests psicológicos y educativos.* Madrid: Síntesis.

Mayén, S. (2009). *Comprensión de medidas de tendencia central en estudiantes mexicanos de educación secundaria y bachillerato.* Tesis Doctoral. Departamento de Didáctica de la Matemática. Universidad de Granada. On line: http://www.ugr.es/~batanero/publicaciones%20index.htm

Mayén, S. (2006). *Comprensión de medidas de posición central en estudiantes mexicanos de Bachillerato.* Memoria de Tercer Ciclo. Departamento de Didáctica de la Matemática. Universidad de Granada.

Morales, P. (1988). *Medición de actitudes en psicología y educación*. San Sebastián: Universidad de Comillas.

Muñiz, J. (1994). *Teoría clásica de los tests.* Madrid: Pirámide.

Santisteban, C. (1990). *Psicometría. Teoría y práctica en la construcción de tests.* Madrid: Norma.

Sax, G. (1989). *Principles of educational and psychological measurement and* evaluation. Belmont, CA: Wadsworth.

SEP (2006). *Programa de estudio, educación secundaria.* Dirección General de Desarrollo Curricular de la Subsecretaría de Educación Básica de la Secretaría de Educación Pública, México.

Thorndike, R. L. (1989). *Psicometría aplicada.* México: Limusa.