ANÁLISIS IMPLICATIVO QUE REFUERZA VALIDEZ Y FIABILIDAD DE UN CUESTIONARIO DE MEDIDAS DE TENDENCIA CENTRAL

Silvia Mayén y Carmen Díaz
CINVESTAV - IPN (México) – Universidad de Huelva (España)
mayazuc@gmail.com, carmen.diaz@dpsi.uhu.es

El Análisis implicativo es un método estadístico que utilizamos para reforzar la validez y fiabilidad de un instrumento de evaluación obtenidas con métodos clásicos, el cual llevamos a cabo sobre las respuestas de un cuestionario que evalúa la comprensión de las medidas de tendencia central y que aplicamos a 643 estudiantes mexicanos de Educación Secundaria y Bachillerato. Nuestro interés se basa en las dificultades que presentan los estudiantes en el estudio de la estadística básica. Obtuvimos también validez de contenido; validez discriminante mediante análisis de diferencia de ejecución en los ítems en los grupos; aproximación a la fiabilidad mediante el cociente Alfa, coeficiente de Theta y Teoría de la generalizabilidad; índice de dificultad de los ítems del cuestionario y una comparación ítem a ítem de los resultados en los distintos grupos de estudiantes. Los resultados confirman que el cuestionario puede ser generalizable a estudiantes de otros contextos y niveles educativos.

PALABRAS CLAVE

Análisis estadístico implicativo, validez, fiabilidad, medidas de tendencia central.

INTRODUCCIÓN

El origen de este estudio es la investigación previa de Cobo (2003) sobre la comprensión de las medidas de tendencia central, que tenía como propósito comparar el nivel de conocimientos de estudiantes españoles que inician y finalizan la educación secundaria obligatoria, de edades comprendidas entre 12 y 16 años, para lo que construyó un cuestionario de dieciséis ítems abiertos relacionados con este tema. En nuestro caso, incluimos alumnos mexicanos que finalizan la educación secundaria y el bachillerato, ya que se preparan para cambiar de nivel, de secundaria a bachillerato y de bachillerato a la universidad. Los ítems de este instrumento presentan situaciones comprensibles para nuestros estudiantes y sus contenidos son semejantes a los programas de estudios mexicanos de ambos niveles. Llevamos a cabo un análisis cuantitativo de evaluación, que consiste en la aplicación del cuestionario a todos los estudiantes. Incluye la validación de dicho cuestionario mediante tres tipos de evidencia de validez: 1) validez de contenido, justificada mediante análisis teórico de los ítems; 2) validez discriminante, mediante análisis de diferencia de ejecución en los ítems en los grupos; y 3) validez de constructo, analizando la estructura de las respuestas mediante análisis clúster e implicativo. Así mismo, se lleva a cabo la aproximación a la fiabilidad mediante el coeficiente Alfa (Martínez Arias, 1995), coeficiente Theta (Barbero, 2003) y Teoría de la Generalizabilidad (Feldt y Brennan, 1991). Abarca también un estudio global de la dificultad de los ítems del cuestionario

y una comparación ítem a ítem de los resultados en ambos grupos de estudiantes. En este trabajo, nos centraremos en el Análisis clúster y Análisis implicativo de dicho estudio. Describimos las muestras de estudiantes, las características del cuestionario y la metodología de su aplicación.

MARCO DE REFERENCIA

Entendemos la *validación de un cuestionario* como un proceso por el cual se aportan evidencias que apoyen la interpretación propuesta de los datos recogidos mediante la prueba (Carmona, 2004) y que está basada tanto en los procesos de respuesta, como en las consecuencias de la evaluación. En nuestro caso, la consecuencia de la evaluación es el diagnóstico de los conocimientos de los estudiantes.

La muestra. Está compuesta por 518 estudiantes mexicanos: 162 del tercer año de Secundaria (14 y 15 años de edad) y 356 del tercer año de bachillerato (17 y 18 años). Incluimos también, una segunda muestra de 125 estudiantes de bachillerato, de un estudio piloto (Mayén, 2006), que tenía como finalidad comprobar que el cuestionario era adecuado para aplicarlo en el contexto mexicano en cuanto a dificultad, contextos de los ítems y con estudiantes de mayor edad. Al contar con una muestra de mayor tamaño, obtenemos una estimación más precisa de la fiabilidad.

El cuestionario. Es un instrumento de medición que por medio de las preguntas planteadas se obtiene una estimación de conocimientos y capacidades de los sujetos a quienes se les aplica, que no son accesibles por simple observación o encuesta (Barbero, 2003). El cuestionario que utilizamos (Mayén, 2009), está orientado a evaluar el significado que los estudiantes mexicanos asignan a las medidas de tendencia central. Fue construido por Cobo (2003) a partir de un análisis de contenido de libros de texto españoles de secundaria y de las directrices curriculares españolas. En nuestro caso también analizamos el currículo mexicano. Contiene los siguientes tipos de elementos: 1) reconocimiento de los campos de problemas; 2) diferentes definiciones de media, mediana y moda; 3) comprensión de propiedades básicas: numéricas, algebraicas y estadísticas; 4) reconocimiento y uso del lenguaje matemático verbal, numérico y gráfico; 5) cálculo y procedimientos de resolución de problemas, algoritmos de cálculo; y 6) argumentaciones. Por tanto, de las respuestas escritas trataremos de inferir el uso (correcto o incorrecto) que los estudiantes de la muestra hacen de los objetos matemáticos mencionados. Incluye un total de 16 ítems, algunos de ellos divididos en subítems (en total 27 subítems) y todos son de respuesta abierta para poder recoger con detalle los razonamientos de los estudiantes.

RESULTADOS

Para el estudio analizamos la *dificultad comparada de los ítems* y el rendimiento total de los estudiantes en la prueba, así como las características psicométricas del cuestionario para completar su validación. Los resultados son los siguientes:

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). Esta validación

se ha hecho mediante el examen sistemático del contenido del test para probar su representatividad y relevancia. Con ello comprobamos que los ítems de nuestro test son relevantes para el uso que se dará a las puntuaciones, y son representativos del contenido que se quiere evaluar, atendiendo a sus características esenciales.

Dificultad comparada de ítems. Para este análisis, categorizamos las respuestas al cuestionario como correctas o incorrectas. La Tabla 1 presenta los *índices de dificultad* de cada ítem, entendiéndose como la proporción de sujetos que aciertan entre todos los que trataron de resolverlo (Muñiz, 1994). Cuanto mayor es este valor, significa que el ítem es más fácil y ha sido respondido correctamente por una mayor proporción de alumnos. Este índice fluctúa entre 0.24 en el ítem 2.3 (cálculo de media de una suma de variables), 0.26 en el ítem 10.1 (cálculo de media, mediana y moda de un conjunto de datos agrupados en intervalos y presentados en una tabla de frecuencias absolutas), y 0.85 en el ítem 8 (estimación de una cantidad desconocida a partir de diversas mediciones en presencia de errores). La mayoría de ítems tiene una dificultad moderada y 24 de los 27 subítems tiene un índice de dificultad de entre 0.3 y 0.7. Se observa que algunos ítems son difíciles de resolver para el global de alumnos.

	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

Tabla 1. Índice de dificultad, Intervalos de Confianza y Credibilidad del 95%

Estimaciones bayesianas. Utilizamos los métodos bayesianos como complemento de los métodos clásicos de inferencia, por algunas razones como la interpretación más intuitiva de los resultados proporcionados (Díaz, 2007) y la posibilidad de considerar la información previa sobre la población en estudio. Por tanto, llevamos a cabo la estimación clásica y bayesiana de los índices de dificultad de los ítems del cuestionario usando la información obtenida por Cobo (2003). De esta forma mejoramos nuestras estimaciones y ofrecemos una interpretación más natural de los intervalos en torno a estas estimaciones (intervalos de credibilidad). En la Tabla 1 aparecen los *intervalos de confianza* (estimación clásica) y los intervalos de credibilidad (estimación bayesiana) para los índices de dificultad de cada uno de los ítems. El primero de ellos ha sido calculado con la fórmula ordinaria de intervalos de confianza de una proporción y tiene una 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*, 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. Para analizar las diferencias que introducen las estimaciones bayesianas, consideramos una distribución inicial uniforme, es decir, supusimos equiprobables todos los valores de la proporción a priori en los diferentes ítems. Por ello, los intervalos de confianza y credibilidad son muy similares.

Análisis de la puntuación total en el cuestionario. Asignamos el valor 1 a cada respuesta, por lo que al sumarlas podríamos obtener una puntuación total de entre 0 y 27 puntos. El número de respuestas correctas osciló entre 5 y 23; el número medio de respuestas correctas es de 13.42 sobre 27, lo que nos da resultados aceptables acercándose a lo esperado, pues el punto medio sería 13.5. La mediana (Mdn=13) indica que la mitad de los estudiantes responden al menos la mitad de los ítems del cuestionario. Todo ello nos asegura unas buenas características del cuestionario, ya que nos permite discriminar una amplia gama de conocimientos, que van desde muy bajos hasta conocimientos altos, encontrándose la mayoría de los estudiantes alrededor del centro, es decir, contestando correctamente a la mitad de las preguntas. El conjunto central de alumnos (50% central) responde entre 10 y 16 preguntas correctamente de un total posible de 27 puntos; no hay sujetos atípicos, lo que es otro indicador de las buenas características del cuestionario. En todo caso hay una gran variabilidad en el número de respuestas correctas, es decir, una comprensión desigual de las medidas de centralización por estos estudiantes.

Análisis de la fiabilidad del cuestionario en nuestra muestra. *Fiabilidad* es la extensión por la cual un experimento, test u otro procedimiento de medida produce los mismos resultados en ensayos repetidos (Cobo, 2003). 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 objeto de estudio. La *fiabilidad* es esta tendencia a la consistencia o precisión del instrumento en la población medida (Bisquerra, 1989). Para calcular la fiabilidad y generalizabilidad usamos todos los datos recogidos, en total 643 estudiantes con el fin de obtener una estimación más precisa de la fiabilidad. Consideramos primero, el

método de *consistencia interna*, que está basado sólo en la aplicación del cuestionario (Díaz, Batanero y Cobo, 2003). Su cálculo se basa en el análisis relativo de la varianza de la puntuación total del cuestionario y de las varianzas de los ítems particulares; el coeficiente que lo mide es el Alfa de Cronbach (Carmines y Zeller, 1979), obteniendo 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, por lo contrario, incluye una gama amplia de conceptos por lo que no es de esperar una fiabilidad excesivamente alta. Seguidamente, calculamos 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 este coeficiente. A partir de los resultados del análisis factorial se calculó el coeficiente Theta de Carmines, (Carmines y Zeller, 1979), el coeficiente alfa presenta un valor bastante alto de 0.726.

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, analizar diferentes fuentes de error en un proceso de medida. El núcleo de esta teoría es considerar diferentes fuentes de error en las puntuaciones observadas (Santisteban, 1990), que pueden ser los mismos sujetos, las preguntas o las condiciones que se aplican. 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) considerando los tamaños de la muestra (27 ítems y 643 alumnos), según si consideramos como fuente de variación los problemas o los alumnos. Obtenemos para la generalizabilidad *respecto a los ítems*, *0.6616,* un valor próximo al coeficiente Alfa (0.662), debido a que en el coeficiente de generalizabilidad respecto a los ítems se considera el número de alumnos fijo y la única fuente de variación se debe a la variabilidad entre ítems. La generalizabilidad *respecto a los alumnos es 0.961,* un valor muy alto que indica una muy alta posibilidad de generalizar nuestros resultados a otros alumnos conservando el mismo cuestionario, bajo la hipótesis conservar las características sociológicas y educativas.

Análisis clúster y análisis estadístico implicativo

Para las interrelaciones entre objetivos de aprendizaje, llevamos a cabo varios análisis multivariantes de las respuestas a los ítems de la prueba. Los resultados se presentan mediante análisis cluster, grafo implicativo y análisis implicativo jerárquico.

Para este análisis utilizamos el software CHIC, *Classification Hierarchical, Implicative et Cohesive* (Couturier y Gras, 2005), que realiza un estudio de aglomeración jerárquica, tomando como medida de similaridad entre ítems el índice de Lerman (Lerman, 1981) 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:

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

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 tenga en cuenta el tamaño de la 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. La medida de similaridad induce un orden parcial en el conjunto de ítems. El primer paso es calcular estos índices, y posteriormente una similaridad entre las clases *A* y *B*; el algoritmo de clasificación jerárquica se construye considerando la mayor proximidad entre elementos de una clase y la mayor distancia entre clases separadas (Tabla 2).

Paso	Nodos que se unen	Similaridad
1	(2.1 2.2)	1
2	((2.1 2.2) 2.3)	1
3	(5.1 5.2)	1
4	(6.1 6.2)	1
5	(12 13)	1
6	(7.2 7.3)	1
7	(15 16)	1
8	((12 13) 14)	1
9	(7.1 (7.2 7.3))	1
10	(9.1 9.2)	0.999999
11	(10.2 10.3)	0.999997
12	(((12 13) 14) (15 16))	0.999996
13	(5.3 (((12 13) 14) (15 16)))	0.99981
14	(10.1 (10.2 10.3))	0.999694
15	((5.3 (((12 13) 14) (15 16))) 11)	0.997586
16	(((2.1 2.2) 2.3) ((5.3 (((12 13) 14) (15 16))) 11))	0.977147
17	(1.1 1.2)	0.960553
18	((5.1 5.2) (9.1 9.2))	0.956171
19	((((2.1 2.2) 2.3) ((5.3 (((12 13) 14) (15 16))) 11)) (10.1 (10.2 10.3)))	0.936159
20	((1.1 1.2) (7.1 (7.2 7.3)))	0.769849
21	(3 4)	0.673794
22	((((2.1 2.2) 2.3) ((5.3 (((12 13) 14) (15 16))) 11)) (10.1 (10.2 10.3))) ((5.1 5.2) (9.1 9.2)))	0.551825
23	((3 4) 8)	0.447176
24	(((1.1 1.2) (7.1 (7.2 7.3))) (((((2.1 2.2) 2.3) ((5.3 (((12 13) 14) (15 16))) 11)) (10.1 (10.2 10.3))) ((5.1 5.2) (9.1 9.2))))	0.342608
25	((((1 1.2) (7.1 (7.2 7.3))) (((((2.1 2.2) 2.3) ((5.3 (((12 13) 14) (15 16))) 11)) (10.1 (10.2 10.3))) ((5.1 5.2) (9.1 9.2)))) ((3 4) 8))	0.0727032
26	(((((1.1 1.2) (7.1 (7.2 7.3))) (((((2.1 2.2) 2.3) ((5.3 (((12 13) 14) (15 16))) 11)) (10.1 (10.2 10.3))) ((5.1 5.2) (9.1 9.2)))) ((3 4) 8))(6.16.2))	0.0171955

Tabla 2. Coeficientes de similaridad en análisis jerárquico según pasos en la clasificación

CHIC proporciona una prueba de significación de las clases obtenidas:

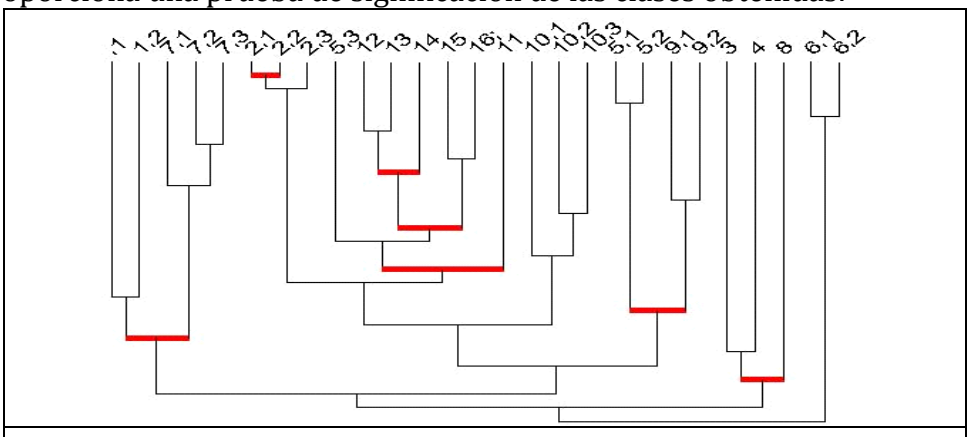

Figura 1. Árbol de similaridad con todas las variables, método clásico, ley binomial

Se calcula considerando la intensidad de la similaridad y el total de sujetos que da una respuesta correcta al conjunto de ítems incluidos en el grupo. El dendograma (Figura 1) muestra los grupos formados de ítems donde los mismos alumnos dan contestaciones similares (correctas o incorrectas). Se trata de conocimientos relacionados entre sí y separados de los otros grupos de ítems. Obtuvimos distintos grupos:

Grupo 1: ítems 1.1 y 1.2 (media como reparto equitativo y determinar una distribución dada la media), que se unen con los ítems 7.1, 7.2 y 7.3 (hallar una distribución dada la media y el efecto del cero sobre el cálculo de la media).

Grupo 2: formado por varios subgrupos, de los cuales solo mencionamos algunos:

Grupo 2.1: ítems 2.1, 2.2 (cálculo de media ponderada) y 2.3 (media de la suma de dos variables). Existe una alta similaridad, pues en la media ponderada fallan muchos estudiantes y en nuestro caso dio un alto índice de errores.

Grupo 2.6: ítems 6.1 y 6.2 (cálculo de mediana en datos ordinales). Relacionados entre sí y completamente separados del resto, por lo que la capacidad de trabajo con datos ordinales no se relaciona con el resto de competencias medidas en el cuestionario.

Grafo implicativo

El análisis clúster considera una medida de asociación simétrica en las variables, es decir, se supone que si un estudiante responde a un ítem también responderá al 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. Para analizar este punto hemos llevado a cabo un *análisis implicativo* entre ítems, que 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*). El software también calcula la significación estadística del índice, que sigue una distribución *N(0.1).* En nuestro caso, las variables aleatorias son las respuestas (correcta- incorrecta) a cada ítem, con lo que podemos calcular un total de $\binom{27}{2}$ índices de implicación entre los 27 ítems del cuestionario.

La relación de implicación entre variables establece un preorden asimétrico dentro del conjunto de ítems, que podemos representar en un grafo ordenado. CHIC calcula los índices de implicación entre todos los pares de variables de un conjunto de datos y proporciona un grafo mostrando todas las implicaciones que son significativas hasta el nivel pedido por el usuario, teniendo en cuenta la intensidad de implicación y el número de sujetos en que se cumple la relación de implicación. La relación de implicación es asimétrica, indicándose el sentido de la implicación por la dirección de la flecha. Por tanto, si estudiamos las relaciones significativas al 99%, la interpretación es que los alumnos resuelven correctamente el ítem (Figura 2). Mencionamos algunas de las relaciones:

- El ítem 6.1 implica al 6.2. El alumno que resuelve correctamente el cálculo de la mediana con datos ordinales en un número impar de valores también lo hará para

un número par de datos. Estos dos ítems están aislados del resto, reafirmando que la competencia con datos ordinales no se relaciona con el resto de los ítems.

- Aparecen aislados los ítems 1.2 (hallar una distribución de datos dada una media) con el 1.1 (media como reparto equitativo). El estudiante que es capaz de formar la distribución es porque comprende este significado de la media.
- El alumno que es capaz de determinar la mediana a partir de un gráfico (ítem 9.1) también podrá hallar la media, ya que el concepto de mediana es más complejo, y por otro lado, también podría 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 distribución representada en el gráfico.

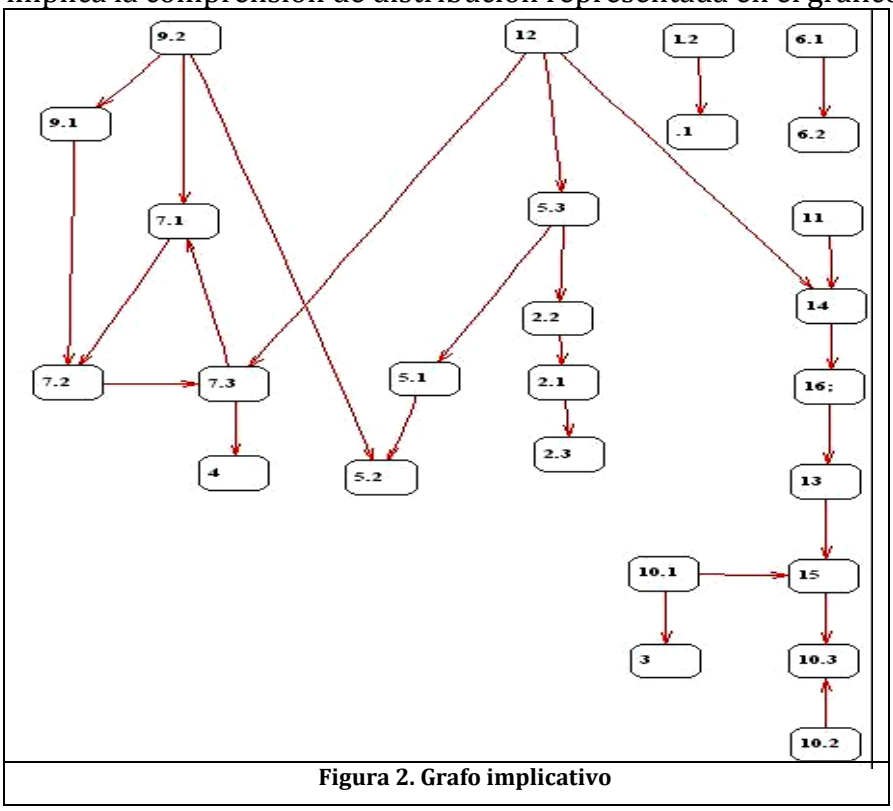

Figura 2. Grafo implicativo

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

Análisis implicativo jerárquico

El grafo implicativo muestra la estructura de interrelaciones y es algo complejo, por lo que se podría dividir el conjunto de ítems en unos pocos grupos interrelacionados entre sí mediante el índice de implicación. Así, llevamos 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. 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 bajo los criterios de la cohesión máxima dentro de cada clase y el mayor grado de implicación entre una clase y otra que es implicada por ella. El estudio concluye con la determinación de una jerarquía implicativa en el conjunto de variables, presentando en la Tabla 4, los coeficientes de cohesión.

Nivel	Nodos que se unen	Cohesión
1	(2.1 2.2)	1
2	((2.1 2.2) 2.3	1
3	(5.1 5.2)	1
4	(6.1 6.2)	1
5	(7.1 7.2)	1
6	((7.1 7.2) 7.3)	1
7	(9.2 9.1)	1
8	(10.2 10.3)	1
9	(16 15)	1
10	(12 13)	1
11	((12 13) 14)	1
12	(1.2 ((7.1 7.2) 7.3))	1
13	(5.3 (5.1 5.2))	0.998
14	(10.1 3)	0.998
15	(((12 13) 14) 11)	0.992
16	((1.2 ((7.1 7.2) 7.3)) 4)	0.986
17	((16 15) .1)	0.977
18	(((2.1 2.2) 2.3) ((1.2 ((7.1 7.2) 7.3)) 4))	0.964
19	((5.3 (5.1 5.2)) (9.2 9.1))	0.88
20	((((12 13) 14) 11) (10.2 10.3))	0.771
21	((10.1 3) ((16 15) 1.1))	0.771

Tabla 4. Coeficientes de similaridad en análisis jerárquico según pasos en la clasificación

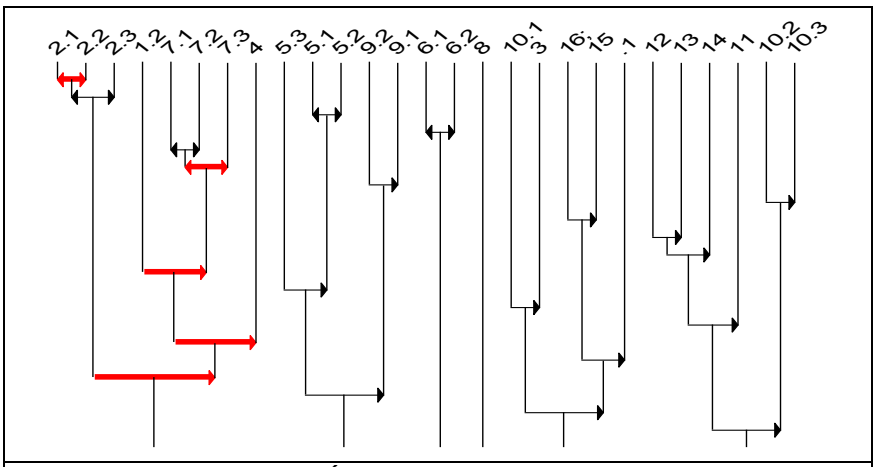

Figura 3. Árbol de cohesión implicativa

En el árbol de cohesión implicativa (Figura 3) se observan cinco grandes grupos:

1. *Cálculo avanzado de la media y comprensión procedimental*. Ítems 2.1, 2.2 (media ponderada) y 2.3 (media de la suma de dos variables), que se implican entre sí; 1.2 (determinar una distribución dada la media), que implica a 7.1, 7.2, 7.3 (hallar una distribución dada la media y el efecto del cero sobre el cálculo de media), que se implican entre sí e implican al ítem 4 (media como valor dentro del recorrido).

2. *Cálculo de la mediana.* Ítem 5.3, que implica a 5.1 y 5.2 (mediana con un número par e impar de datos), que se implican entre sí e implican al 9.2 (mediana a partir de un gráfico), que a su vez, implica al 9.1 (cálculo de media partir de un gráfico).
3. *Mediana en datos ordinales.* Ítems 6.1 y 6.2 (cálculo de mediana en datos ordinales), que se implican entre sí y completamente separados del resto.
4. *Comprensión conceptual de la media.* Ítem 10.1 (cálculo de media en una tabla), que implica al ítem 3 (suma de desviaciones a la media), que implican a los ítems 16 (interpretación de promedios en gráfico conjunto de dos distribuciones) y el 15 (cambio de escala), y que implican al 1.1 (media como reparto equitativo).
5. *Representante de un conjunto de datos.* Ítem 12 (efecto del valor atípico sobre los promedios), que implica al ítem 13 (mejor representante en distribución no simétrica), que se unen a su vez con el 14 (distribución bimodal) y luego con el 11 (significado de media y mediana e interpretar mediana y moda en un gráfico), y 10.2, que implica al 10.3 (cálculo de media, mediana y moda a partir de una tabla).
6. El ítem 8, *media como mejor estimación,* aparece separado del resto.

CONCLUSIONES

Los resultados del estudio de análisis cluster e implicativo muestran la estructura compleja del cuestionario e indican que la comprensión de ciertos elementos de significado de las medidas de tendencia central no se relaciona con los de otros. Observamos cómo los grupos que se relacionan con la comprensión conceptual (de definiciones o de propiedades) aparecen separados de los relacionados con aspectos procedimentales y también aparecen como separados los grupos de ítems relacionados con la mediana (grupos dos y tres) de los relacionados con la media. De este modo se sugiere que la comprensión de la mediana y de la media no está relacionada en estos estudiantes. Se observa también que el tratamiento de mediana con datos ordinales no es habitual para los estudiantes, lo que indica mayor énfasis en la enseñanza. Otra aportación de este estudio es la validación del cuestionario, pues hemos argumentado, al igual que hizo Cobo (2003) su validez de contenido, pero se ha añadido el estudio de la validez discriminante y de constructo que Cobo (2003) no llevó a cabo. Hemos repetido también los cálculos de índices de fiabilidad y generalizabilidad, pero ahora se completan con el coeficiente Theta basado en el análisis factorial, más adecuado para cuestionarios multidimensionales como este. También se proporcionan estimaciones bayesianas de los índices de dificultad del cuestionario, muy apropiadas para situaciones con información previa, como la nuestra. Respecto a la validez de constructo, se ha repetido el análisis clúster que realizó Cobo (2003), añadiéndose ahora el análisis implicativo para confirmar la estructura de las respuestas al cuestionario.

REFERENCIAS

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

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

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

Carmona, J. (2004). Una revisión de las evidencias de fiabilidad y validez de los cuestionarios de actitudes y ansiedad hacia la estadística. *Statistics Education Research Journal*, *3* (1). http://www.stat.auckland.ac.nz/ ~iase/serj/.

Cobo, B. (2003). Significado de las medidas de posición central para los estudiantes de secundaria. Tesis de doctorado. Granada: 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). Francia: Universidad de Lille.

Díaz, C. (2007). Viabilidad de la inferencia bayesiana en el análisis de datos en psicología. Tesis de doctorado. Granada: 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.* New York: John Wiley.

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

Lerman, I.C. (1981). *Classification et analyse ordinale des données.* París : Dunod.

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 de doctorado. Granada: Universidad de Granada.

Mayén, S. (2006). Comprensión de medidas de posición central en estudiantes mexicanos de Bachillerato. Memoria de Tercer Ciclo. Granada: Universidad de Granada.

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