FIABILIDAD Y GENERALIZABILIDAD. APLICACIONES EN EVALUACIÓN EDUCATIVA[1]

Carmen Díaz[(1)] , Carmen Batanero[(1)] y Belén Cobo[(2)]

[(1)] Universidad de Granada, [(2)]Instituto de Enseñanza Secundaria de Huétor Vega, Granada

En este trabajo analizamos los conceptos de fiabilidad y generalizabilidad, reflexionando sobre su utilidad en la evaluación educativa. Describimos los diferentes coeficientes, indicando las situaciones en que podrían aplicarse y qué información proporciona cada uno de ellos, analizando también los factores que influyen en su valor. Mostramos con un ejemplo los pasos a seguir para el cálculo de la fiabilidad mediante el método de consistencia interna y los coeficientes de generalizabilidad, utilizando los datos recogidos sobre un cuestionario de comprensión de promedios en dos muestras de alumnos (14 años; n=168 y 17 años; n=144).

1. INTRODUCCIÓN

Según Thorndike (1989), el proceso de medida en educación y psicología se propone ligar ciertos conceptos abstractos a indicadores empíricos. En el caso de la evaluación educativa, intentamos relacionar los conocimientos de los alumnos sobre un concepto, sus capacidades y estrategias en la resolución de problemas y tareas o sus actitudes con sus respuestas a los ítems de una prueba o de un cuestionario. El análisis de datos se hace sobre las respuestas, ya que son observables. El interés teórico, sin embargo, es el concepto subyacente (conocimientos, actitudes, etc.) que no podemos observar directamente, pero que tratamos de inferir a partir de las respuestas.

Cuando la relación entre los indicadores empíricos (respuestas) y los conceptos subyacenteS es fuerte, el análisis de los indicadores nos permite hacer inferencias útiles sobre los conceptos teóricos y evaluar nuestras hipótesis previas sobre los mismos, así como tomar decisiones adecuadas sobre la acción didáctica. Para medir esta relación se han definido diversos conceptos, como validez, fiabilidad o generalizabilidad. Puesto que, cuando evaluamos a un alumno, podemos plantearle una infinidad de posibles preguntas sobre el mismo tema, e incluso con las mismas preguntas, las repuestas del alumno pueden variar dependiendo de su atención, cansancio u otros factores, debemos reconocer un carácter aleatorio a los resultados de la evaluación y, en general, de una investigación educativa.

Las fuentes de error en este proceso pueden ser de naturaleza determinista o aleatoria y afectan de diversa forma a nuestras conclusiones y decisiones. Los *sesgos*, de naturaleza determinista, aunque de magnitud desconocida, se derivan de nuestros procedimientos, tanto en la selección de la muestra, como en la elaboración de los instrumentos y en la toma o análisis de los datos. Por ejemplo, en un estudio de evaluación de conocimientos puede producirse un sesgo si la prueba de evaluación es demasiado sencilla (por lo cual los alumnos más brillantes no mostrarán toda su capacidad) o no contempla adecuadamente la variable que queremos medir (por ejemplo, si queremos evaluar la capacidad de resolución de problemas, en general, y sólo proponemos problemas aritméticos sencillos). Los sesgos suelen afectar al valor de

[1] *Números,* 54, 3 – 21.

las variables siempre en la misma dirección, no disminuyen al aumentar el tamaño de la muestra y pueden ser evitados cambiando los métodos utilizados. La ausencia de sesgo se conoce como **validez.** No entraremos en este trabajo en la discusión de los diversos tipos de validez y su comprobación. Remitimos al lector al texto de Muñiz (1994) para más detalle.

En este trabajo nos preocupamos del control de los errores *aleatorios* que son debidos a la variabilidad de la muestra, del contexto o del material experimental. Suelen afectar al valor de las variables, unas veces en exceso y otras en defecto, por lo que, al aumentar el tamaño de la muestra se disminuye su efecto. Hablamos de precisión o **fiabilidad**. Recientemente, el concepto de fiabilidad se ha ampliado dando paso al de *generalizabilidad* que es menos conocido y utilizado en educación. Un segundo objetivo de este trabajo es comparar los conceptos de fiabilidad y generalizabilidad mostrando su cálculo y utilidad e indicando qué información proporcionan los respectivos coeficientes. Para facilitar la exposición usaremos los datos obtenidos en un cuestionario de comprensión de promedios, que se han obtenido en dos muestras de alumnos (14 años; n=168 y 17 años; n=144).

2. DESCRIPCIÓN DEL CUESTIONARIO Y MUESTRA

El cuestionario que usaremos como ejemplo ha sido elaborado como parte de una investigación sobre la comprensión de los promedios y está dirigido a alumnos que cursan la Educación Secundaria Obligatoria en España (13-17 años). Está formado por 16 ítems, con un total de 25 subítems que pretenden evaluar:

a) Comprensión de las propiedades básicas y definiciones de media, mediana y moda.

b) Reconocimiento de representaciones verbales, simbólicas y gráficas.

c) Cálculo y procedimientos de resolución de problemas. Comprensión de los algoritmos de cálculo frente a su aplicación automática.

d) Reconocimiento de problemas de promedio y capacidad de argumentación.

Los ítems se han tomado de diversas investigaciones previas (Cai, 1995; Carvalho, 1998, Cobo, 1998; Mevarech, 1983; Pollatsek, Lima, y Well, 1981; Strauss, y Bichler, 1988; Tormo, 1993 y Watson, y Moritz, 2000), que se centran en la comprensión de puntos específicos, como el cálculo o una cierta propiedad, y donde, generalmente, se presentaban con formato de opciones múltiples. En nuestro caso se trata de evaluar globalmente la comprensión de los puntos que hemos indicado con un sólo cuestionario y también permitir las respuestas abiertas. También se ha tenido en cuenta las orientaciones curriculares en España (M.E.C. 1992) y los contenidos en los libros de texto españoles de educación secundaria (Cobo, 2001). En el anexo hemos incluído los cuatro primeros ítems, y en la Tabla 1 se reflejan los conocimientos evaluados por cada uno de ellos. Usaremos estos cuatro ítems que son suficientes para ejemplificar los conceptos que vamos a introducir, aunque los coeficientes de fiabilidad y generalizabilidad que presentamos se calcularon sobre el cuestionario completo.

El cuestionario fue administrado a una muestra de 312 alumnos de 14 y 17 años procedentes de cinco centros educativos y de los cursos 1º y 4º de ESO. El grupo de 1º de ESO se componía de 168 alumnos y el grupo de 4º de ESO de 144 alumnos. Los grupos estaban igualados en género.

Tabla 1. Conocimientos evaluados en cuatro ítems de un cuestionario

Conocimientos evaluados	Item						
	1.1	1.2	2.1	2.2	2.3	3	4
Definición de media	X	X	X	X	X	X	X
La operación "promediar" no conserva el conjunto numérico	X						
La media podría no coincidir con ninguno de los datos	X						
Media de la suma de dos variables					X		
La media es conmutativa					X		
La media no es asociativa					X		
La media es un valor representativo	X						
Suma de desviaciones a la media						X	X
Hallar un valor representativo			X	X			
Efectuar un reparto equitativo	X					X	
Predecir un valor probable	X						
Cálculo de la media		X	X	X			
Cálculo de medias ponderadas			X	X			
Invertir el algoritmo de la media		X					X
Encontrar una distribución de media dada		X					X

Para comprender mejor el concepto de fiabilidad, que se refiere al total del cuestrionario, es interesante analizar primero la idea de *índice de dificultad,* que se refiere a cada ítem aislado. Siguiendo a Muñiz (1994), entenderemos por índice de dificultad de un ítem (ID) la proporción de sujetos que lo resuelven correctamente. Los índices de dificultad y desviaciones típicas de los cuatro ítems se presentan en la tabla 2. Observamos en dicha tabla que los ítems no son homogéneos ni en tanto a su dificultad ni en cuanto a su variabilidad. No son tampoco homogéneos respecto a su contenido, midiendo una variedad de conocimientos, que se refleja en la Tabla 1 y que es mucho mayor en el cuestionario completo. Resaltamos este hecho porque un cuestionario muy homogéneo suele tener una alta fiabilidad, pero puede estar sesgado (carecer de validez) para evaluar la comprensión de un concepto o una parte de las matemáticas, ya que al ser las preguntas muy homogéneas, evaluaremos sólo una parte del contenido. En nuestro caso, buscamos a propósito preguntas heterogéneas, como se refleja en la Tabla 1 o en el enunciado de los ítems que aparecen en anexo, con la finalidad de que los alumnos pudieran expresar una amplia gama de razonamientos y habilidades. En otros trabajos hemos estudiado con detalle el cuestionario y los resultados obtenidos en cada ítem (Cobo y Batanero, 2001; Batanero, Cobo y Díaz, en prensa).

Tabla 2. Indice de dificultad y desviación típica de los ítems por curso

	Primer Curso n=168		Cuarto Curso n=144	
	Media	Desv. típ.	Media	Desv. típ.
P1.1	.63	.48	.69	.46
P1.2	.27	.44	,37	,48
P2.1	.14	,34	,34	,48
P2.2	.12	,32	,38	,49
P2.3	.36	,48	,33	,47
P3	.46	,50	,49	,50
P4	.45	,50	,66	,48

3. FIABILIDAD Y SU ESTIMACIÓN

Para cualquier instrumento de medida es necesario conocer la precisión de las medidas que nos proporciona, ya que esta precisión se puede usar para extender los resultados de la muestra particular a una población más general (por ejemplo, calculando intervalos de confianza). Un instrumento de medida se considerará *fiable* si las medidas que se obtienen a partir de él no contienen errores o los errores son suficientemente pequeños.

En el modelo lineal clásico (Muñiz, 1994) la puntación empírica X obtenida para un sujeto en el total de una prueba es la suma de dos componentes: la puntuación verdadera (V) del sujeto en ese test o prueba y el error de medida (e). El modelo más sencillo relacionando estas dos puntuaciones hace las siguientes hipótesis:

$$X=V+e$$

$$E(X) =V$$

Donde *E(X)* es la esperanza matemática o media de la variable aleatoria *X*. En realidad este supuesto define la puntuación verdadera, ya que no tenemos modo de medirla. Suponemos que la puntuación verdadera del sujeto es la puntuación media que se obtendría si aplicásemos repetida e indefinidamente la prueba al sujeto.

Se supone también que *e* es una variable aleatoria que sigue una distribución normal de media cero (es decir, los errores positivos y negativos se compensan) y que no está correlacionada con X (por tanto el error es independiente de la puntuación empírica obtenida.

Se entiende por *fiabilidad* de un test, un cuestionario u otro instrumento de medida la estabilidad de las puntuaciones que proporciona si se administra en repetidas ocasiones al mismo grupo de personas. Un supuesto implícito en el estudio de la fiabilidad es la estabilidad de la variable que se pretende medir. La medida siempre produce un cierto error aleatorio, pero dos medidas del mismo fenómeno sobre un mismo individuo suelen ser consistentes. La fiabilidad es esta tendencia a la consistencia o precisión del instrumento en la población medida (Bisquerra, 1989).

Coeficiente de fiabilidad y sus diversas estimaciones

El *coeficiente de fiabilidad* es un indicador de la fiabilidad teórica de las puntuaciones observadas, en el sentido de proporcionar un valor numérico que indica el *grado de confianza* que podíamos tener en dichas puntuaciones como estimadores de las puntuaciones verdaderas de los sujetos. Se define como la correlación entre las puntuaciones *X* y *X'* obtenidas por el sujeto en un test cuando se le pasa dos veces sucesivas. Muñiz (1994) indica que, si no hubiese errores de medida, las puntuaciones coincidirían y la correlación sería perfecta, por lo que este coeficiente sería igual a 1.

Este coeficiente de fiabilidad mide, por tanto, la extensión por la que nuestro instrumento de medida está afectado por los errores aleatorios y es un valor teórico que debe ser estimado por algún procedimiento empírico, a través de las respuestas de un grupo de sujetos a un conjunto de items. Hay diversos procedimientos para el cálculo del estimador del coeficiente de fiabilidad, que resumimos a continuación.

Test- retest

En este método se administra el mismo test dos veces a las mismas personas y se calcula el coeficiente de correlación entre las puntuaciones obtenidas en las dos ocasiones. Intenta medir el porcentaje de variabilidad debido a las fuentes que contribuyen a que un sujeto tenga diferente puntuación en aplicaciones repetidas de la "misma prueba": temporales, ambientales, estado, sentimientos... y da una medida de la estabilidad del rasgo en las personas durante el periodo de tiempo dado. Por ello se le denomina *coeficiente de estabilidad*. El uso principal de este coeficiente es decidir si se van a utilizar unas puntuaciones que fueron obtenidas para un grupo de sujetos en una prueba anterior o si se ha de repetir la prueba en el momento actual. Sin embargo se plantean los siguientes problemas:

- El intervalo de tiempo que se deja entre las dos administraciones del test es difícil de determinar: tiene que ser suficiente para que no recuerden la tarea pero no demasiado amplio para que no se den cambios en los sujetos (aprendizaje, maduración...).
- A veces solo podemos medir la variable una sola vez en un instante fijo de tiempo. Por ejemplo, si son alumnos que ya han abandonado el centro escolar.
- Una baja correlacion test/retest puede ser debida no a la falta de fiabilidad sino a un cambio en el concepto subyacente, sobre todo cuanto más tiempo haya pasado entre las pruebas. Es el caso de que se haya producido una maduración en los sujetos.
- Puede haber una reactividad del instrumento, que puede influir sobre el mismo concepto en el individuo. Esto ocurre, por ejemplo, en los test de actitudes.
- La memoria de las personas sobre sus contestaciones en la primera prueba pueden influenciar las segundas, sobre todo cuando el intervalo de tiempo transcurrido es demasiado corto.

A pesar de estos inconvenientes, este coeficiente es muy util en algunos casos, como cuando se mide la fiabilidad de un sistema de códificacion o de transcripcion de los datos.

Formas paralelas

Este método de estimación requiere construir dos formas paralelas del test que son administradas al mismo grupo de sujetos dejando entre ambas administraciones un periodo de tiempo breve para evitar cambios en los sujetos, aunque suficiente para evitar la fatiga. Se eliminan así muchos de los inconvenientes que hemos señalado para el test /retest.

Al calcular la correlación entre las puntuaciones de ambas administraciones, se obtiene directamente una estimación del coeficiente de fiabilidad del test. Las componentes de variabilidad que se desean medir por este método son las debidas a fluctuaciones de disposición de unos items respecto a otros o ligeros cambios en los ítems. Nos indica la estabilidad de variaciones de la misma prueba o de items que han sido diseñados para medir las mismas funciones. El coeficiente obtenido se denomina *coeficiente de equivalencia*.

Se supone que los tests paralelos miden la misma capacidad o rasgo y de la misma manera. En cada una de las formas paralelas, se supone que los sujetos poseen la misma

puntuación verdadera y las varianzas error para ambos es la misma. Por tanto, ambos tests tendrán la misma media y la misma varianza observada. La dificultad de este método de estimar la fiabilidad es la construcción de dos formas estrictamente paralelas del test y, a pesar de que existen desarrollos teóricos que demuestran que bastaría con formas "razonablemente" paralelas, cualquier coeficiente calculado mediante este procedimiento estará influido pro la falta de similitud entre las formas. Es aconsejable que la mitad de los sujetos utilicen la primera vez la forma A y la segunda la B y luego al contrario, para estudiar el posible efecto del orden de administración.

Test- retest con formas paralelas:

Es una combinación de los dos procedimientos anteriores. Se administra una forma del test, se deja transcurrir un periodo de tiempo semejante al del procedimiento Test-retest y se administra la forma paralela del test a la misma muestra de sujetos. El coeficiente que se obtiene se denomina *coeficiente de estabilidad y equivalencia*.

Los procedimientos anteriores se basan en evaluar dos veces a cada sujeto. A continuación señalamos los procedimientos basados en una sola administración del test. Con estos métodos se pretende determinar el grado de consistencia de las respuestas de los sujetos, ver si los sujetos responden de forma consistente a lo largo del conjunto de items. Por esta razón se denominan estos coeficientes se denominan *de consistencia interna*.

Dos mitades del test

Una vez pasado el test, se divide en dos mitades y se calcula la puntuación del sujeto en cada mitad. Al dividir el test se debe conseguir que las dos mitades sean equivalentes, algo que en ocasiones puede resultar difícil. Se pueden dividir aleatoriamente, según la dificultad o según el contenido y la dificultad (el procedimiento más aconsejable). Se calcula la correlación entre las puntuaciones obtenidas en las dos mitades del test, con un procedimiento de corrección que puede ser el de Spearman-Brown:

siendo r el coeficiente de correlación empírico entre las dos mitades del test.

Covariación entre los items del test:

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. También es una cota inferior de la que se obtendría por el método de la prueba repetida si se comparase el test dado y otro cualquiera paralelo de igual cantidad de ítems (Carmines y Zeller, 1979). Hay dos métodos de cálculo de la fiabilidad basados en la covariación entre los items del test:

- Alfa de Cronbach: refleja el grado en el que covarían los items que constituyen el test.

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n} s_j^2}{S_n^2}\right)$$

- Kuder- Richardson: sólo se aplica a items dicotómicos. Su fórmula es la siguiente:

siendo *n* el número de ítems, S^2 la varianza de la puntuación total, s_j^2 las varianzas de cada ítem, que en el caso de ser dicotómicos son iguales al producto , siendo el índice de dificultad del ítem.

La interpretación de la fiabilidad de consistencia interna está estrechamente ligada con las anteriores. El coeficiente alfa para un test de N items es igual al valor medio de todos los que se obtendrían con el método de las dos mitades si se utilizasen todas las combinaciones de items. Es la estimación de una *fiabilidad en el acto*. Se acerca la puntuación de una persona a la que se hubiese obtenido si tuviésemos un instrumento perfecto de medición.

Interpretación de los diversos coeficientes

La preferencia por uno de los tipos de estimación de la fiabilidad depende tambien del tipo de prueba. En primer lugar hay que distinguir entre pruebas homogéneas y no homogéneas. Una prueba muy homogénea mide la misma habilidad en todas sus partes, es un asunto de coherencia interna. Si la prueba es heterogénea no hay que esperar un índice de consistencia interna muy alto. Las partes de una prueba semejante tampoco se correlacionarán positivamente si se eligen los items al azar. Por tanto, la única estimación con sentido de la fiabilidad en las pruebas hetereogéneas es la repetición de la prueba.

Por otro lado hay que distinguir entre las pruebas de velocidad y de potencia. En esta última hay suficiente tiempo para todos los items y se espera que cada examinando trate de responder todos los ítems. En una prueba de velocidad seria inapropiado utilizar la fiabilidad de dos mitades o de coherencia y tampoco sería válido la de repetición de la prueba, si hay efecto de aprendizaje. Habría que utilizar una forma paralela del test.

Factores que afectan a la fiabilidad

Los factores más comunes son los siguientes:

- La fiabilidad de la persona que corrige la prueba o que transcribe los datos. Cuanto más alto sea el acuerdo de dos personas que corrigen las mismas pruebas mayor será la fiabilidad.
- Variabilidad del grupo al que se pasa la prueba. Cuanto más variable sea el grupo, mayor será la fiabilidad. Consideremos, por ejemplo que la varianza fuese cero. Ello querría decir que todos los sujetos tendrían la misma puntuación. Es fácil deducir,

de la fórmula del coeficiente de correlación, que en tal caso, la correlación sería cero, y por ello también lo sería el coeficiente de fiabilidad.

- Número de ítems de la prueba. Un test es más fiable, a mayor número de ítems. Ello se debe en parte a que un mayor número de ítems aumenta la varianza de la prueba y también por la fórmula del coeficiente de fiabilidad, que es función de este número.
- Índices de dificultad de los ítems. Si los ítems son extremadamente sencillos, todos los alumnos responderán todos los ítems, e igual ocurre si son extremadamente difíciles. Esto afecta a la variabilidad reduciéndola.

Relación entre fiabilidad y validez

Hay una creencia generalizada de que la validez de una prueba está relacionada directamente con su fiabilidad. Para interpretar esta regla hay que diferenciar entre pruebas homogéneas y heterogéneas. Las primeras tienen una alta fiabilidad de consistencia interna, pero pueden no tener validez para medir ciertos rasgos al dejar de incluir aspectos importantes. En las segundas puede ocurrir lo contrario.

La fiabilidad y validez pueden ser, por tanto, incompatibles y hay que llegar a un compromiso. La fiabilidad muy alta requiere ítems muy correlacionados entre sí y la validez muy alta ítems poco correlacionados. Además, es deseable una gama de dificultades, con objeto de disponer de una gama graduada para medir las diversas capacidades de las personas.

4. ESTIMACIÓN DE LA FIABILIDAD EN UN CUESTIONARIO SOBRE PROMEDIOS

Para aclarar los conceptos anteriores y mostrar la forma en que pueden aplicarse, analizaremos como ejemplo el cálculo de la fiabilidad en nuestro cuestionario. Al decidir, cuál método sería factible, entre los diversos métodos de estimar la fiabilidad de una escala, hemos tomado, en nuestro caso, el método de *consistencia interna,* ya que solo disponíamos de una administración del test. Elegimos un método basado en la correlación entre los items del test, usando el coeficiente Alfa de Cronbach, al tratarse de ítems no dicotómicos.

Es importante destacar la necesidad de realizar el cálculo para cada grupo por separado por varias razones: los alumnos de cada grupo son de diferente edad, los índices de dificultad son diferentes en cada grupo y también la variabilidad de la puntuación total, factores estos que afectan a la fiabilidad de la escala. Por otro lado, el grupo de Cuarto de ESO realizó el test completo (un total de 25 ítems), mientras que los alumnos de 1º curso sólo contestaron a un total de 17 subítems. Como información complementaria hemos calculado el coeficiente de fiabilidad conjunto para los 16 ítems comunes a los dos grupos. Puesto que la respuesta a cada subítem se puntúa separadamente, es necesario tratar cada subítem como una variable en el cálculo de la fiabilidad.

El cálculo se ha realizado mediante el programa SPSS, subprocedimiento Análisis de fiabilidad, dentro de la opción Escalas. Este procedimiento calcula los estadísticos. descriptivos para cada variable y para la escala, estadísticos de resumen comparando los elementos, correlaciones y covarianzas inter-elementos, diversas estimaciones de la

fiabilidad y otros estadísticos. Proporciona los diferentes coeficientes de fiabilidad que hemos descrito.

Fiabilidad para el Primer Curso

En la Tabla 3 presentamos los resultados que proporciona el programa SPSS para los cuatro primeros ítems en la muestra de alumnos de 1º de ESO (en nuestro caso se realizó el cáculo con los 17 items, aunque no presentamos la tabla completa en este artículo). Se obtuvo un valor $\alpha = 0.7076$ para el coeficiente de Cronbach. Aunque este valor sugiere una correlación fuerte, no es excesivamente elevado cuando se interpreta como índice de fiabilidad. La razón para ello es que el cuestionario incluye items de contenidos diversos (conceptos, cálculo e interpretación de la media, la mediana y la moda), por lo que no es un cuestionario muy homogéneo y por tanto la correlación entre los items del cuestionario no es muy elevada (podemos verlo en la columna de correlación del ítem con el total). Aún así el coeficiente era lo suficientemente elevado para el propósito de la investigación para la que se diseñó el cuestionario.

También SPSS proporciona para cada ítem información sobre la forma en que afecta a la fiabilidad global. Vemos por ejemplo, que eliminar el ítem P3 implica aumentar la fiabilidad de la escala. Este ítem 3 (suma de desviaciones a la media) correlaciona negativamente con el total de la prueba. Aunque el valor negativo que aparece es muy pequeño, -0.07 merece una reflexión, puesto que la varianza sin el ítem es la más elevada de toda la tabla, 0.73. Estos dos valores indican que evalúa componentes diferenciados respecto al resto de los ítems del cuestionario. En nuestro caso, estábamos interesados en evaluar concretamente la comprensión de una propiedad específica, por lo que, incluso cuando el ítem disminuya la fiabilidad total de la escala se decidió conservarlo.

Tabla 3. Resultados del análisis de fiabilidad para primer curso

Ítem	Media sin el ítem	Varianza sin el ítem	Correlación con el total	Alfa sin el ítem
P1.1	5,15	8,93	,28	,69
P1.2	5,51	8,50	,50	,67
P2.1	5,64	8,95	,44	,68
P2.2	5,66	8,99	,46	,68
P2.3	5,42	8,70	,37	,68
P3	5,32	9,98	-,07	,73
P4	5,33	8,75	,33	,68

Alfa = 0,7076

Por lo que respecta a los otros ítems en el ejemplo, la correlación con el total de la prueba presenta valores que van desde 0.28 para el 1.1 (definición de media), hasta el máximo de 0.50 que corresponde al 1.2 (dada la media, encontrar una distribución con dicha media) y que es por tanto un item que discrimina los alumnos con mejor y peor comprensión. Es el que más contribuye a la fiabilidad, entre los presentados en el ejemplo.

El resto de los valores (Media sin el ítem, Varianza sin el ítem y Alfa sin el ítem), permanecen muy estables para las diferentes cuestiones. Ello señala una contribución homogénea de cada ítem a la puntuación en la escala, lo que refuerza la elección hecha

de esta escala para la investigación. Estas medias y varianzas se refieren al valor total teórico (17 ítems) y por tanto indican por un lado, dificultad global del cuestionario (media en torno a 6 respuestas correctas) así como varianza reducida (en comparación con la media prevista). Estos dos factores de nuevo explican el valor moderado del coeficiente de fiabilidad.

De forma análoga se realizaron los analisis de fiabilidad para el cuarto curso y el analisis de las preguntas comunes a los dos grupos con el global de la muestra. En la Tabla 4 presentamos los coeficientes obtenidos.

4. COEFICIENTES DE GENERALIZABILIDAD

La teoría de la generalizabilidad extiende la teoría clásica de la medición, según 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 el considerar diferentes fuentes de error en las puntuaciones observadas, que pueden ser los mismos sujetos, las preguntas o las condiciones que se aplican.

El coeficiente de generalizabilidad se define con el cociente (1),

es decir como cociente entre la varianza de las puntuaciones verdaderas de la prueba y

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

la varianza observada, que es suma de la varianza verdadera más la varianza debida al error aleatorio. Según Thorndike (1989), la varianza de error depende de cómo definimos el universo de puntuaciones verdaderas y en el análisis de generalizabilidad se consideran ciertas fuentes como parte de la varianza de error en unas condiciones y otras fuentes en otras. En nuestro caso diferenciaremos dos fuentes para el error aleatorio y calcularemos, por tanto, dos coeficientes de generalizabilidad: la generalizabilidad a otros alumnos de la misma prueba y la generalizabilidad si pasáramos otros problemas similares a los incluidos en la prueba a los mismos alumnos.

Para realizar este cálculo, hemos obtenido, en primer lugar a partir del análisis de escalas del programa SPSS y del modelo de estimación de Dunn y Clarck (1987) para el análisis de varianza de medida repetida, los componentes de la varianza de las puntuaciones observadas. El análisis de escalas de SPSS proporciona, si se le pide, una Tabla de Análisis de varianza de medidas repetidas (Tabla 4). De esta tabla obtenemos los cuadrados medios entre sujetos, entre los diferentes ítems y residual, así como sus grados de libertad.

Tabla 3. Analisis de varianza de medidas repetidas para Primer curso

Fuente de variación	Suma cuadrados	G.L.	Cuadrado medio	F	Prob
Entre sujetos	98.25	167	.58		
Intra sujetos	542.94	2688	.20		
Entre items	83.26	16	5.20	30.25	.0000
Residual	459.67	2672	.17		
Total	641.19	2855	.22		
Media total	.34				

$CM_S = 0.58$ que es un estimador de $b\sigma^2_S + \sigma^2_R$, siendo b el número de ítems

$CM_I = 5.20$ que es un estimador de $a\sigma^2_I + \sigma^2_R$ siendo a el número de sujetos

$CM_R = 0.17$ que es un estimador de σ^2_R

De donde, despejando obtenemos las siguientes estimaciones:

Varianza dentro de los sujetos $\sigma^2_s = 0.0246$

Varianza dentro de los ítems $\sigma^2_i = 0.0299$

Varianza residual $\sigma^2_r = 0.172$

Sustituyendo ahora estos componentes de varianza en la fórmula (1) y teniendo en cuenta los tamaños de muestra (17 ítems y 168 alumnos), según si consideramos como fuente de variación los problemas o los alumnos, obtenemos estimaciones que analizamos a continuación.

Generalizabilidad respecto a otros ítems:

Cuando queremos generalizar a otros ítems, conservando fijos los sujetos, consideramos la varianza dentro de los sujetos como varianza de la puntuación verdadera. La varianza de error en cada ítem será la varianza residual, dividida por el número de ítems. La fórmula (1) del coeficiente de generalizabilidad se transforma en este caso en la expresión (2)

(2)

En nuestro ejemplo, obtenemos un valor próximo al del coeficiente Alfa, lo cual es lógico, puesto que el coeficiente de generalizabilidad a otros ítems coincide con él, ya que se considera los alumnos fijos y la única fuente de variación es la debida a variabilidad entre ítems. Las pequeñas diferencias son debidas a redondeos en los cálculos. Este coeficiente mide la generalizabilidad de nuestros resultados si a los mismos alumnos les pasáramos otra prueba del mismo número de ítems, variando el enunciado de los mismos.

Generalizabilidad a otros alumnos:

Cuando queremos generalizar a otros sujetos, conservando fijos los items, consideramos la varianza dentro de los ítems como varianza de la puntuación verdadera. La varianza de error en cada sujeto será la varianza residual, dividida por el número de sujetos. La fórmula (1) del coeficiente de generalizabilidad se transforma en este caso en la expresión (3):

(3)

Obtenemos en nuestro ejemplo un valor muy alto, para la generalizabilidad a otros alumnos de la misma prueba, lo que indica una muy alta posibilidad de generalizar nuestros resultados a otros alumnos, conservando el mismo cuestionario. Por supuesto, en la hipótesis de que se conserven las características sociológicas y educativas de los alumnos a los que se pasa la prueba.

Como resumen, presentamos en la tabla 4 los distintos coficientes calculados. Observamos que cualquiera de los coeficientes cambia al cambiar el grupo de sujetos a que se aplica la prueba. Es por ello que siempre es necesario el cálculo de la fiabilidad, aún cuando usemos un cuestionario construído por otro autor, quien ya hubiese calculado estos coeficientes.

Tabla 4. Resultados comparados de los coeficientes de fiabilidad y generalizabilidad

	Fiabilidad (Alfa)	Generalizabilidad (Items)	Generalizabilidad (Alumnos)
Primer Curso n=168	0,7076	0,710	0,9670
Cuarto Curso n=144	0,7607	0,76067	0,9661
Grupo combinado n=312	0,7879	0,78789	0,9804

Una segunda consecuencia es que la teoría de la generalizabilidad amplía con ventaja el cálculo clásico de la fiabilidad, puesto que permite calcular al menos dos coeficientes uno de los cuales (generalizabilidad a otros ítems) coincide con el coeficiente de fiabilidad. Observamos también la importancia de aumentar en lo posible el número de ítems así como el tamaño de la muestra. Ambos factores mejoran la fiabilidad de los resultados, como vemos en nuestro ejemplo.

Finalmente es importante destacar como una prueba aparentemente sólo moderadamente fiable (no muy alta generalizabilidad a otros ítems, como es el caso de nuestro cuestionario) puede ser altamente generalizable en cuanto los resultados obtenidos se quieren generalizar a otros alumnos.

CONCLUSIONES

En la investigación en educación necesitamos con frecuencia aplicar o construir cuestionarios para evaluar diversas variables no observables, tales como conocimientos o actitudes. La dificultad de llevar a cabo inferencias correctas de los resultados obtenidos a las variables que queremos medir se deduce de la multitud de factores que pueden influir en la variabilidad de las respuestas a la prueba, así como de la infinitas posibilidades de variación que tenemos al construir una misma prueba de evaluación.

Los coeficientes de fiabilidad y generalizabilidad tratan de dar una medida objetiva de la estabilidad de las puntuaciones obtenidas frente a variaciones aleatorias. En este trabajo hemos tratado de mostrar la aplicabilidad de estos coeficientes y la forma en que pueden calcularse a partir de SPSS. Pensamos que esta información es útil a los investigadores, quienes con frecuencia aplican estos cálculos en forma mecánica.

Es necesario también desmitificar el significado de estos coeficientes, cuya interpretación siempre dependerá de los fines de la evaluación. En un contexto puramente

educativo, lo que nos interesa primordialmente son los sujetos concretos que son evaluados. Nuestro interés es proporcionar una información lo más fiable posible sobre los alumnos de un curso y sobre cada uno de ellos y no estamos, por tanto, interesados en extender nuestros resultados a unos alumnos diferentes. El coeficiente de fiabilidad proporciona en estos casos una información suficiente. Cuando el interés es realizar conjeturas o predicciones sobre las capacidades de otros alumnos más allá de la muestra utilizada, sería importante el cálculo de los coeficientes de generalizabilidad.

REFERENCIAS

Batanero, C., Cobo, B. y Díaz, C. (en prensa). Assessing secondary school students' understanding of averages. *III European Mathematics Education Conference.*

Cobo, B. y Batanero, C. (2001). Razonamientos aritméticos en problemas de promedios. En J. M. Cardeñoso y otros (Eds.), *Investigación en el Aula de Matemáticas*. Atención a la diversidad (pp. 149-157). Granada: Sociedad Thales. IBSN: 84-699-6874-2.

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

Cai, J. (1995). Beyond the computational algorithm. Students'unsderstanding of the arithmetic average concept. En L. Meira (Ed.). *Proceedings of the 19th PME Conference* (v.3, pp. 144-151). Universidade Federal de Pernambuco, Recife, Brasil.

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

Carvalho, C. (1998). Tarefas estadísticas e estratégias de resposta. Comunicación presentada en el *VI Encuentro en Educación Matemática de la Sociedad Portuguesa de Ciencias de la Educación*. Castelo de Vide, Portugal.

Cobo, B. (1998). *Estadísticos de orden en la enseñanza secundaria.* Memoria de Tercer Ciclo. Departamento de Didáctica de la Matemática. Universidad de Granada.

Cobo, B. (2001). Problemas y algoritmos relacionados con la media en los libros de texto de secundaria. *Jornadas Europeas de Enseñanza y Difusión de la Estadística.* Palma de Mallorca: Instituto Balear de Estadística.

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

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

M.E.C. (1992). *Decretos de Enseñanza Secundaria Obligatoria*. Madrid: Ministerio de Educación y Ciencia.

Mevarech, Z.R. (1983). A deep structure model of students'statistical misconceptions. Educational Studies in Mathematics, 14, 415-429.

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

Pollatsek, A., Lima, S. y Well, A. D. (1981). Concept or Computation: Students'understanding of the mean. *Educational Studies in Mathematics*, 12, 191-204.

Strauss, S. y Bichler, E. (1988). The development of children's concepts of the arithmetic average. Journal for Research in Mathematics Education, 19 (1), 64-80.

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

Tormo, C. (1993). *Estudio sobre cuatro propiedades de la media aritmética en alumnos de 12 a 15 años*. Memoria de Tercer Ciclo. Universidad de Valencia.

Watson, J. M. y Moritz, J. B. (2000). The longitudinal development of understanding of average. *Mathematical Thinking and Learning*, v1 (2/3), 11-50.

Anexo. Cuestionario

Ítem 1. Un periódico dice que el número medio de hijos por familia en Andalucía es 1.2 hijos por familia.

1. Explícanos qué significa para ti esta frase.
2. Se han elegido 10 familias andaluzas y el número medio de hijos entre las 10 familias es 1.2 hijos por familia. Los García tienen 4 hijos y los Pérez tienen 1 hijo, ¿cuántos hijos podrían tener las otras 8 familias para que la media de hijos en las diez familias sea 1.2? Justifica tu respuesta.

Ítem 2. Maria y Pedro dedican una media de 8 horas cada fin de semana a hacer deporte. Otros 8 estudiantes dedican cada semana una media de 4 horas a hacer deporte.

1. ¿Cuál es el número medio de horas que hacen deporte cada fin de semana los 10 estudiantes?.
2. María y Pedro dedican además 1 hora cada fin de semana a escuchar música y los otros 8 estudiantes, 3 horas. ¿Cuál es el número medio de horas que escuchan música los 10 estudiantes?
3. ¿Cuál sería el número medio de horas que estos 10 estudiantes dedican, cada fin de semana, entre las dos actividades?

Ítem 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?

Ítem 4. Tenemos seis números y el más grande es el 5. Sumamos estos números y dividimos la suma por seis. El resultado es 4. ¿Te parece posible? ¿Por qué?