Revista Investigación Educativa - Vol. 8 - n.º 16 - 1990 (P. 483-491)

LA TEORÍA DE LA GENERALIZABILIDAD COMO BASE PARA LA EVALUACIÓN Y TOMA DE DECISIONES

por
L. Buendía Eisman
M. Pegalajar Moral
Universidad de Granada

I. INTRODUCCIÓN

En todos estos coeficientes de fiabilidad descritos en la literatura existente, tanto los índices de acuerdo como los coeficientes de correlación, hemos encontrado diferentes limitaciones. Los coeficientes de fiabilidad basados en el acuerdo intra e interobservadores, confunden el error aleatorio del sujeto con las diferencias intraobservadores e interobservadores. Los índices basados en el coeficiente de correlación, confunden igualmente el error aleatorio del sujeto con las diferencias entre las subdivisiones o formas; e incluso el índice de constancia (puntuaciones de un observador acerca de dos sesiones), confunde los errores de medida con los cambios reales que se producen en la conducta del sujeto en los dos momentos de la observación (Blanco, 1986).

Estos métodos, por tanto, no consideran simultáneamente las diferentes fuentes de error. Es preciso una teoría multivariada que tenga en cuenta, además de las diferencias individuales entre sujetos, otras fuentes de variación (observadores, ocasiones, formas diferentes) y que permita integrar cada una de esas fuentes de error en una estructura global. Esto se considera en la teoría de la generalizabilidad desarrollada por Cronbach, Gleser, Nanda y Rajaratnam (1972) y cuya aplicación pasamos a ver a continuación.

2. APLICACIÓN DE LA TEORÍA DE LA GENERALIZABILIDAD EN LA EVALUACIÓN

El uso de la observación directa para obtener datos sobre los profesores, alumnos, o clima de clase en general, es común en la Investigación Educativa y en la evaluación de contextos.

Los investigadores, o personas responsables de la obtención de los datos, desean que sus medidas sean fiables, de modo que se puedan tomar confiadamente decisiones, en base a los datos obtenidos. Sin embargo con frecuencia, las decisiones sobre la frecuencia y duración de los períodos de observación, el nº de evaluadores, el tipo de cuestiones a observar, etc., son elegidos más por intuición o por la experiencia que por análisis empíricos.

Los coeficientes de fiabilidad raramente son utilizados para estas decisiones y, normalmente, son incluidos post-hoc como coeficientes de correlación o coeficientes intraclase, obtenidos con alguno de los procedimientos anteriormente citados. Los coeficientes que incluyen múltiples fuentes de variación, aunque poco utilizados en la investigación educativa, proporcionan una formulación que es aplicable a medidas observacionales, atribuyendo la variación en estas medidas a múltiples fuentes, con lo que habrá que calcular diferentes coeficientes de fiabilidad para la mayoría de los contextos observacionales. Estas fuentes de variación son denominadas facetas, en los estudios G, son análogas a las facetas en el análisis de varianza. Un *componente de varianza es estimado por cada faceta* y un coeficiente de generalizabilidad se estima utilizando la fórmula siguiente:

$$(\text{Rho})^2 = \frac{\text{Varianza de la puntuación universo estimada}}{\text{Varianza de la puntuación observada estimada}}$$

Crombach y colaboradores, utilizan el término Rho^2 para designar el coeficiente de generalizabilidad cuando Rho^2 proporciona un índice de la garantía de las medidas para diferenciar entre niveles de facetas especificadas y generalizar de unos niveles a otros. Los componentes de varianza son utilizados para estimar, la varianza de las puntuaciones universo y la varianza de la puntuación observada, para las facetas que nos interesan.

El empleo de los coeficientes de generalizabilidad para conocer la fiabilidad de los datos recogidos sobre variables de proceso en clase, supone un campo de evidente aplicación en Investigación Educativa.

A través de ellos se pueden responder los siguientes tipos de cuestiones: ¿Qué facetas y cuantos niveles de cada una es necesario muestrear para obtener estimaciones fiables de la conducta docente? ¿Pueden usarse diferentes observadores para observar a diferentes maestros o deben los mismos observadores observar a todos los maestros? ¿Puede un profesor (que toma la decisión en una clase) generalizar un resultado de su conducta a todas las asignaturas o la generalización estará restringida a la conducta docente en, por ejemplo, matemáticas?

Las dos primeras preguntas afectan al uso de diseños cruzados frente a diseños anidados; mientras que la tercera afecta a si las facetas deben ser tratadas como aleatorias o fijas. A continuación presentamos algunos ejemplos para ilustrar cada uno de estos diseños.

3. DISEÑOS PARA LOS ESTUDIOS «G»

Utilizando la teoría de la generalizabilidad con un diseño de dos facetas, obtendremos los mismos resultados que con un coeficiente α (Cronbach, 1951) o los coeficientes intraclase obtenidos utilizando los procedimientos de Hoyt (1941). Mientras que los procedimientos de Hoyt y Cronbach no son aplicables a situaciones que impliquen más de dos facetas, la teoría de la generalizabilidad puede ser aplicada a contextos que impliquen fuentes múltiples de variación. En el trabajo presente, la formulación G, se introduce en un diseño de dos facetas y luego, se amplía para incorporar tres fuentes de variación. Se tratarán, en varios ejemplos que incluyen datos recogidos con el instrumento para evaluación de la actuación o movimiento del profesor (TPAI), de Capie, Johnson, Anderson, Ellett & Okey (1979) y aplicado por K. G. Tobin (1980) y por nosotros, en la investigación actualmente en vias de realización.

Diseño de dos facetas

En una investigación diseñada para estimar la generalizabilidad de las medidas del rendimiento del profesor, seis evaluadores evaluaron cada uno a cuatro profesores con el instrumento TPAI. Se obtuvo una sola medida del rendimiento por cada profesor, sumando, las puntuaciones de competencia relacionadas con las habilidades interpersonales, y los procedimientos de la clase. Este diseño puede ser representado cómo un diseño completamente cruzado, con los evaluadores y profesores, representando las facetas en un análisis de varianza de dos vías. La fórmula para estimar los componentes de varianza para estos diseños se ofrecen en la Tabla 1.

TABLA 1. ESTIMACIÓN DE LOS COMPONENTES DE VARIANZA EN UN DISEÑO CRUZADO DE DOS FACETAS

Fuentes de variación	nivel	EMS	Componentes de varianza
Profesores (P)	j		$\hat{\sigma}^2_e + \hat{\sigma}^2_{pr} + K\hat{\sigma}^2_p ; \hat{\sigma}^2_p = \frac{(MS_p - MS_{res})}{K}$
Evaluadores (R)	K		$\hat{\sigma}^2_e + \hat{\sigma}^2_{pr} + j\hat{\sigma}^2_r ; \hat{\sigma}^2_r = \frac{(MSr - MS_{res})}{j}$
Residual			$\hat{\sigma}^2_e + \hat{\sigma}^2_{pr} ; \hat{\sigma}^2_{res} = MS$

El coeficiente de generalizabilidad para diferenciar los profesores se calcula con la fórmula siguiente. El componente de varianza atribuible al error se divide por K, número de evaluadores, puesto que las medidas que nos interesan son evaluaciones medias del rendimiento del profesor basadas en K evaluadores.

$$Rho_p^2 = \frac{\hat{\sigma}_p^2}{\sigma_p^2 + 1/K(\hat{\sigma}_{res}^2)}$$

El análisis de varianza que se puede aplicar a este ejemplo se ofrece en la tabla 2. Utilizando las estimaciones de componentes de varianza de la tabla 2, se obtuvo un coeficiente de 0'99. La diferenciación de los profesores resultó bastante satisfactoria utilizando éste procedimiento.

TABLA 2. ANÁLISIS DE VARIANZA PARA UN DISEÑO CRUZADO DE DOS FACETAS

Fuentes de variación	d f	M S	nivel	componentes de varianza
Maestros P	3	1856'92	4	301'89
Evaluadores R	5	101'42	6	13'96
Residual	15	45'58		45'58

Un procedimiento similar se utilizó para diferenciar entre evaluadores utilizando la fórmula siguiente:

$$Rho_r^2 = \frac{\hat{\sigma}_r^2}{\hat{\sigma}_r^2 + 1/j(\hat{\sigma}_{res}^2)}$$

Se obtuvo un coeficiente G de 0'55, que es ciertamente menor que el valor de 0'80 que se suele utilizar como valor de comparación (Cardiner, Tourneur y Allal 1976). Este coeficiente indica que las medidas de rendimiento del profesor no son fiables para diferenciar entre evaluadores y no deberían ser utilizadas para orientarnos a la hora de tomar decisiones sobre los evaluadores. El diseño de un estudio G determinará el nº de coeficientes que pueden ser calculados. Como norma general, se puede obtener más información de un diseño completamente cruzado ya que los coeficientes pueden ser estimados para todas las facetas. Para introducir el concepto de anidamiento, consideremos el ejemplo siguiente que es una modificación del primer diseño.

Seis profesores son evaluados cada uno por tres evaluadores; cada evaluador evalua solamente a un profesor, de modo que cada profesor es evaluado por un conjunto diferente de evaluadores. La fórmula para estimar los componentes de varianza para éste ejemplo está contenida en la tabla 3, y los datos de ANOVA se ofrecen en la tabla 4.

TABLA 3. ESTIMACIÓN DE LOS COMPONENTES DE VARIANZA PARA UN DISEÑO ANIDADO DE DOS FACETAS

Fuentes de variación	nivel	EMS	Componentes de varianza
Profesores (P)	j	$\hat{\sigma}^2_e + \hat{\sigma}^2_{r(p)} + \hat{\sigma}^2_p$	$\hat{\sigma}^2_p = \frac{(MS_p - MS_{res})}{K}$
Residual	K	$\hat{\sigma}^2_e + \hat{\sigma}^2_{r(p)}$	$\hat{\sigma}^2_{res} = MS_{res}$

TABLA 4. ANÁLISIS DE VARIANZA

Fuentes de variación	df	MS	Nivel	Componentes de varianza
Profesores (P)	5	596'13	6	193'84
Residual	12	14'61		14'61

Puesto que cada evaluador sólo evalua a un profesor, la varianza atribuible a las diferencias entre los evaluadores se confunde con la varianza debida al error por la interacción entre profeso y error. Como consecuencia, no se puede estimar un coeficiente de generalizabilidad para diferenciar entre los evaluadores.

El coeficiente de generalizabilidad para diferenciar entre profesores se ofrece a continuación.

$$Rho^2_p = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + 1/K(\sigma^2_{res})}$$

El Rho cuadrado en éste caso es de 0'98. El rendimiento medio del profesor, obtenido a través de las puntuaciones de tres evaluadores ha resultado fiable para diferenciar profesores.

Un Diseño cruzado de tres facetas

Si consideramos las competencias de la enseñanza como una fuente de varianza, introduciríamos una tercera faceta en el diseño. Los procedimientos utilizados para calcular los coeficientes, en los primeros dos ejemplos, son fácilmente extendibles cuando se incorpora una tercera faceta. En estos casos, los maestros y evaluadores son considerados como muestras aleatorias procedentes de un universo de facetas y evaluadores respectivamente. En términos del modelo de análisis de varianza estas facetas son consideradas como efectos aleatorios. Sin embargo, las competencias pueden ser consideradas como una faceta fija, si no son consideradas representativas de un número infinito de competencias similares.

La inclusión de una faceta fija en el diseño puede afectar la magnitud de los coeficientes de generalizabilidad. Los términos de interacción, para el aspecto de diferenciación con un efecto fijo, se incluyen en el numerador de la ecuación para estimar el coeficiente de generabilidad. Las implicaciones de un efecto fijo, son solamente para la generalización. Pero en lo concerniente a la diferenciación de una faceta, no hay diferencias, si ésta es aleatoria o fija. Según Cardinet y al. (1973 p. 13) «La fijación o no fijación de las condiciones de los aspectos de generalizabilidad tiene importantes consecuencias sobre el valor de los coeficientes G. No hemos encontrado consecuencias similares para definir de un modo u otro las condiciones de las facetas de diferenciación».

En los apartados siguientes trataremos los coeficientes de generalizabilidad para un diseño completamente cruzado, en el que seis evaluadores evaluan cada uno a cuatro profesores sobre diez competencias de enseñanza. Se consideran cuatro casos

TABLA 5 . COMPONENTES DE VARIANZA PARA UN DISEÑO CRUZADO DE TRES FACETAS

Fuentes de variación	niveles	EMS	Componentes de varianza
Profesores (P)	j	$\hat{\sigma}^2_e + \hat{\sigma}^2_{prc} + i\hat{\sigma}^2_{pr} + K\hat{\sigma}^2_{pc} + Ki\hat{\sigma}^2_p$	; $\hat{\sigma}^2_p = \frac{MS_p + MS_{pr} + MS_{pc} + MS_{res}}{Ki}$
Evaluadores (R)	K	$\hat{\sigma}^2_e + \hat{\sigma}^2_{prc} + j\hat{\sigma}^2_{rc} + i\hat{\sigma}^2_{pr} + iji\hat{\sigma}^2_r$	; $\hat{\sigma}^2_r = \frac{MS_r + MS_{rc} + MS_{pc} + MS_{res}}{ij}$
Competencias (C)	i	$\hat{\sigma}^2_e + \hat{\sigma}^2_{prc} + j\hat{\sigma}^2_{rc} + K\hat{\sigma}^2_{pc} + jK\hat{\sigma}^2_c$	; $\hat{\sigma}^2_c = \frac{MS_c + MS_{rc} + MS_{pc} + MS_{res}}{jK}$
PXR		$\hat{\sigma}^2_e + \hat{\sigma}^2_{prc} + i\hat{\sigma}^2_{pr}$	; $\hat{\sigma}^2_{pr} = \frac{MS_{pr} + MS_{res}}{i}$
PXC		$\hat{\sigma}^2_e + \hat{\sigma}^2_{prc} + K\hat{\sigma}^2_{pc}$	; $\hat{\sigma}^2_{pc} = \frac{MS_{pc} + MS_{res}}{K}$
RXC		$\hat{\sigma}^2_e + \hat{\sigma}^2_{prc} + j\hat{\sigma}^2_{rc}$	; $\hat{\sigma}^2_{rc} = \frac{MS_{rc} + MS_{res}}{j}$
Residual		$\hat{\sigma}^2_c + \hat{\sigma}^2_{pr}$	; $\hat{\sigma}^2_{res} = MS_{res}$

TABLA 6. ANÁLISIS DE VARIANZA

Fuentes de variación	df	MS	Nivel	Componentes de varianza	
Profesores	3	137'3087	j = 4	2'0054	
Evaluadores	5	9'8576		K = 6	0'0146
Competencias	9	277'0730	i = 10	11'1120	
Profesores x evaluadores	15	8'6951			0'7180
Profesores x competencias	27	9'8053			1'3817
Evaluadores x competencias	45	2'0940			0'1448
Residual	131	1'5150			1'5150

en estos diseños completamente cruzados y consideramos las competencias como una faceta fija, en cada uno de los ejemplos que se presentan.

La tabla 5, ofrece la fórmula utilizada para estimar los componentes de varianza para estos diseños. La tabla 6 contiene las estimaciones de componentes de varianza que se pueden aplicar a los diseños cruzados de tres facetas.

Caso 1. Diferenciar profesores y generalizar a través de evaluadores y competencias

Tal y como fue el caso de los ejemplos anteriores los valores medios de las medidas son las fuentes de diferenciación. Como consecuencia las estimaciones de la varianza se basan en el número de casos utilizados para determinar la media relevante.

$$Rho_p^2 = \frac{\hat{\sigma}_p^2 + 1/i\ \hat{\sigma}_{pc}^2}{\hat{\sigma}_p^2 + 1/i\ \hat{\sigma}_{pc}^2 + 1/K\ \hat{\sigma}_{rp}^2 + 1/iK\ \hat{\sigma}_{res}^2}$$

En éste caso el coeficiente de generalizabilidad es 0,94. Así que los datos del rendimiento del profesor son fiables para diferenciar profesores y generalizar a través de evaluadores y competencias.

Caso 2. Diferenciar evaluadores y generalizar a través de profesores y competencias

Una fórmula similar proporciona una estimación de la capacidad para diferenciar evaluadores y generalizar a través de profesores y competencias.

$$Rho_r^2 = \frac{\sigma_r^2 + 1/j\ \sigma_{rc}^2}{\sigma_r^2 + 1/i\ \hat{\sigma}_{rc}^2 + 1/j\ \hat{\sigma}_{rt}^2 + 1/ij\ \hat{\sigma}_{res}^2}$$

En éste caso se obtuvo un coeficiente de 0,06. Estos datos no son fiables en un contexto en el que hace falta diferenciar a los evaluadores en base a unas puntuaciones medias a través de profesores y competencias.

Caso 3. Diferenciar competencias y generalizar a través de maestros y evaluadores

Si la cuestión en una investigación es determinar si existen diferencias entre competencias, el coeficiente de generalizabilidad apropiado se calcula utilizando la fórmula siguiente. En éste caso las medidas podrían ser puntuaciones medias a través de todos los profesores y evaluadores.

$$Rho^2_c = \frac{\hat{\sigma}^2_c}{\hat{\sigma}^2_e + 1/K\ \hat{\sigma}^2_{rc} + 1/j\ \hat{\sigma}^2_{pe} + 1/jK\ \hat{\sigma}^2_{res}}$$

El valor obtenido para el Rho cuadrado en éste ejemplo es de 0,96. Claramente las medidas obtenidas podrían ser utilizadas para diferenciar competencias y generalizar a todos los evaluadores y profesores.

Caso 4. Variación debida a la interacción entre facetas

El diseño de tres facetas contiene cuatro términos de interacción. La interacción de tres vías se confunde con el error y no puede ser estimada en éste diseño. Una interacción maestro por competencia es posible y sería de esperar en el ejemplo que expondremos a continuación. En contraste, una interacción evaluador por profesor o una interacción evaluador por competencia también son posibles, pero deberían ser pequeñas y atribuibles al error. En éste caso, el coeficiente G debería ser bajo. Un coeficiente elevado supondría la necesidad de cambios en los procedimientos de recogida de datos. Así pues, se pueden obtener sólo dos coeficientes G, para el ejemplo que estamos considerando. El coeficiente, evaluador por profesor, no puede ser calculado porque se confundiría el término del error con la interacción de tres vías.

Los datos en la tabla 7, indican la naturaleza lógica de la interacción profesor por competencia y la naturaleza aleatoria de la interacción evaluador por competencia. (Tobin, 1980).

TABLA 7

Fuentes de variación	fórmula	2
Profesor por competencia	$Rho^2_{pc} = \frac{\hat{\sigma}^2_{pc}}{\hat{\sigma}^2_{pc} + 1/K\ \hat{\sigma}^2_{res}}$	0'85
Evaluador por competencia	$Rho^2_{rc} = \frac{\hat{\sigma}^2_{rc}}{\hat{\sigma}^2_{rc} + 1/j\ \hat{\sigma}^2_{res}}$	0'28

Evaluador por profesor. No puede calcularse con este diseño.

Aunque en este trabajo nos hemos limitado a diseños sencillos, el modelo se puede extender fácilmente incorporando facetas adicionales y diseños complejos de anidación. Para un tratamiento detallado de algunos diseños vease Cronbach y al. (1972).

La teoría de la generalizabilidad proporciona un marco de trabajo para incorporar las fuentes múltiples de variación para la estimación de los coeficientes de fiabilidad de las medidas que van a ser utilizadas en contextos específicos. La formulación G resulta especialmente aplicable a las medidas educacionales que a menudo se ven afectadas por un número identificable de fuentes de varianza. Debido a la dependencia contextual de los coeficientes G habría que tomar precauciones para identificar las fuentes de varianza y para especificar con precisión el modelo estadístico que permita realizar estimaciones no sesgadas de los componentes de varianza.

La importancia de los coeficientes de fiabilidad está en la planificación de las investigaciones y en la interpretación de los datos.

Además, a causa de las relaciones entre fiabilidad y probabilidad de cometer errores tipo II la magnitud de los coeficientes G es importante para la interpretación de resultados, cuando no se han podido confirmar las relaciones hipotéticas.

NOTAS BIBLIOGRÁFICAS

ANGUERA, M. T.: Metodología observional. En J. Arnau, M.T. Anguera y J. Gómez. *Metodologías de Investigación en Psicología* Murcia: Servicio de Publicaciones de la Universidad de Murcia.

BLANCO, A. (1986): *Generalizabilidad de la observación de la conducta.* Trabajo inédito no publicado. Barcelona: Universidad de Barcelona.

BRENNAM, R. and KANE, M. (1979): *Generalizability theory: A review.* New Directions for Testing and measurement, 4, 33-51.

CAPIE, W. JOHNSON, C.E.; ANDERSON, S. J.; ELLET, C. D. and OKEY, J. R. (1979): *Teacher performance Assessment Instruments.* Athens: Teacher Assessment Project, University of Georgia.

CRONBACH, L. J.; GLESER, G. C.; NANDA, H. y RAJARATNAM, N. (1972): *The dependability of generalizability for scores and profiles.* New York: Wiley.

CRONBACH, L. J. (1951): *Coefficient alpha and the internal structure of tests.* Psichometrika 16, 297-334.

HOYT, C. (1941): *Test reliability estimated by the analysis of variance.* Psychometrika, 6, 153-160.

SHAVELSON, R. J. and WEBB, N. M. (1981): *Generalizability theory. 1973-1980.* British Journal of matematical and statistical Psychology. 34. 133-166.

SHAVELSON, R. J. and WEBB, N. M. (1989): *Generalizability theory.* American Psychologdist. Vol 44. 6 922-932.

TOBIN, K. J. (1980): *Applications of Generalizability theory to Classroom Process Measures.* Annual Meeting of the Nai. Ass. for Research in Scien. Teach.