El problema de la fiabilidad en la evaluación de la eficacia docente en la universidad: una alternativa metodológica

por Isabel PASCUAL GÓMEZ y José Luis GAVIRIA SOTO
Universidad Antonio de Nebrija
Universidad Complutense de Madrid

Introducción

A mediados de la década de los veinte se inicia en Estados Unidos la investigación sobre la evaluación de la calidad docente del profesorado universitario aplicando cuestionarios a los alumnos. Desde entonces esta metodología evaluativa ha sido ampliamente utilizada e investigada en los ámbitos universitarios nacionales e internacionales (Cohen 1980; Marsh 1987; Feldman 1996; Fernández 1992; Tejedor 1993; Villa 1993; Apodaka 1999).

La investigación sobre la evaluación de la eficacia docente se ha centrado principalmente en la construcción de instrumentos de evaluación docente que permitiesen conocer las características de un docente eficaz y en el análisis de la *supuesta* estructura dimensional subyacente, con el objetivo último de ofrecer al profesor *feedback* sobre su eficacia, y que éste repercutiese positivamente en su acción docente.

La opinión subjetiva del alumno, recogida a través de estos instrumentos y obtenida de las percepciones sobre la eficacia docente de un profesor, se ha valorado e interpretado bajo el prisma de una métrica objetiva. Se ha convertido la opinión de los alumnos en un hecho o conocimiento, y se ha utilizado una métrica común y estable en el tiempo para interpretarla, validándose su fiabilidad a través del estadístico Alpha de Cronbach.

Sin embargo numerosos han sido los investigadores que han criticado esta forma de evaluación (Cruse 1987; Weissberg 1993; Haskell 1997; Sproule 2000). Las críticas realizadas cuestionan la función diagnóstica y formativa de este tipo de evaluación y desaprueban la tendencia de los responsables políticos y educativos, a utilizar los resultados obtenidos de las encuestas de evaluación como indicadores de calidad en los procesos de selección o promoción laboral del profesorado.

La desconfianza existente sobre la idoneidad de este método de evaluación obedece a diversas razones. A las dificultades de carácter teórico o de indefinición conceptual del constructo *eficacia docente* hay que añadir las dificultades de medición. Autores como Sproule (2000) consideran completamente inadecuado reducir la eficacia docente a "*números sin significado*", y consideran necesario mejorar la fiabilidad y validez de estos procedimientos de valoración de la calidad docente.

Si se entiende la fiabilidad como precisión y ausencia de error, es necesario que los diseños metodológicos utilizados en este tipo de procedimientos profundicen en el análisis de la variabilidad y de los diferentes tipos de errores que pueden producirse en estas evaluaciones.

La verdadera puntuación en eficacia docente de un profesor está sesgada por los errores de medida propios de cualquier procedimiento de medición pero además, por el hecho de que su *eficacia* se obtiene a través de percepciones, no de indicadores objetivos y cuantitativos.

Es necesario, por lo tanto, comprobar la variabilidad de los diversos rasgos que intervienen en esta circunstancia concreta, tanto los relacionados con el alumno como con los condicionantes ambientales. Los condicionantes personales están relacionados con la subjetividad de la opinión del alumno. Los condicionantes ambientales están relacionados con la estructura de los datos de los que se obtiene la evaluación del profesor.

En este sentido, los datos tienen una estructura jerárquica o anidada. Si se ignora la posible ausencia de independencia entre los diversos niveles de datos, característica de este tipo de estructuras anidadas, y se simplifica en exceso el análisis de las fuentes de variabilidad, se pueden cometer graves errores en la interpretación de los resultados.

Partiendo de esta premisa en este artículo se mostrará cómo varían los resultados obtenidos en los índices de fiabilidad de un cuestionario de evaluación docente cuando estos se abordan desde la perspectiva de los modelos jerárquicos lineales, especialmente adaptados a este tipo de estructuras.

Objetivo

El objetivo de este trabajo es desagregar el índice de fiabilidad en sus componentes relativos a la fiabilidad del instrumento respecto a los alumnos, y de los alumnos respecto al profesor mediante una adaptación metodológica de este procedimiento a los modelos jerárquicos lineales.

Metodología general del estudio

Para llevar a cabo la comparación de los dos procedimientos de determinación de la fiabilidad se optó por aplicar ambos tipos de métodos a un cuestionario de evaluación docente de una universidad privada, la Universidad Antonio de Nebrija. El cuestionario de evaluación necesario para éste procedimiento se obtuvo tras la revisión y consulta de los instrumentos de evaluación utilizados en otras universidades españolas como la Universidad

de Valencia, la Universidad del País Vasco, Universidad Complutense de Madrid y la Universidad Carlos III de Madrid entre otras.

El instrumento final es una escala de elección múltiple de 1 a 7 que contiene 20 ítems. Los ítems se agrupan en torno a 4 factores: *Competencia del profesor* (ítems sobre claridad expositiva del profesor, dominio de la asignatura, etc.), *Capacidad de interacción del profesor* (ítems sobre la capacidad de dialogo del profesor o accesibilidad), *Medios y Materiales*, y *Cumplimiento formal* (horario, tutorías, programa). Algunos de los ítems se excluyeron del procedimiento de validación por entender que no estaban relacionados con la eficacia docente del profesor.

El estudio de validez del instrumento se llevó a cabo a través de dos procedimientos:

a) análisis de los ítems y de su relación con la calidad docente.

b) análisis de la dimensionalidad del cuestionario o identificación de los ítems o variables más relacionados entre sí, y su significación común.

La comprobación de la fiabilidad del instrumento, entendida ésta como consistencia o unanimidad de los alumnos a la hora de evaluar, y estabilidad en el tiempo, se obtuvo a través del coeficiente de fiabilidad de Cronbach.

Datos

El procedimiento de validación final de este instrumento, tras un procedimiento piloto inicial se realizó analizando los datos obtenidos de las cinco Facultades y de la Escuela Politécnica Superior que componen la Universidad.

Los datos de la muestra pertenecen a las evaluaciones semestrales realizadas por los alumnos, de forma anónima, durante dos cursos académicos: el curso 97-98 y el curso 98-99. En los cuadros siguientes se describen las poblaciones de alumnos y profesores evaluados.

TABLA 1: *Población de alumnos encuestados*

Facultad	**Derecho**	**Empresas**	**Filología**	**Ingeniería**	**Comunicac.**	**Total**
97-98 1er semestre	150	936	222	752	985	3045
97-98 2do semestre	188	562	152	670	776	2348
98-99 1er semestre	303	829	163	1200	1379	3874
98-99 2do semestre	308	408		635	997	2348
Total						11.615

TABLA 2: *Población de profesores evaluados*

97-98 1er semestre	159
97-98 2do semestre	170
98-99 1er semestre	235
98-99 2do semestre	174
Total	738

Las técnicas implicadas en el proceso de validación fueron:

a) Análisis psicométrico y descriptivo de los ítems.

b) Análisis de validez o análisis de las correlaciones entre los ítems y la correlación ítem-criterio.

c) Análisis factorial o comprobación de la estructura dimensional del instrumento.

d) Análisis de fiabilidad o cálculo del Alpha de Cronbach.

a) El análisis psicométrico de los ítems del instrumento reflejó medias altas en la mayoría de los ítems, índices de dispersión bajos, indicando homogeneidad en las conductas evaluadas y alta correlación de los ítems con el ítem criterio. El ítem criterio se definió como aquel ítem que medía la satisfacción general del alumno con su profesor.

b) Los altos índices de correlación de Spearman obtenidos mostraron que las conductas descritas por los ítems del instrumento eran relacionadas por el alumno con lo que este entendía por enseñanza eficaz, excepto en el caso del ítem *dificultad de la asignatura*. Y por lo tanto se puede afirmar que los índices de validez del instrumento son aceptables como puede observarse en la Tabla 3.

TABLA 3: *Correlaciones entre los ítems del cuestionario y la variable criterio satisfacción global*

Variable	Correlación
Me gustaría cursar otra asignatura con este profesor	0,839
Consigue que los alumnos se interesen por la asignatura	0,749
Responde las dudas con claridad	0,735
Explica con claridad	0,730
Resalta y deja claros los aspectos más importantes de la materia	0,699
Intenta mantener la atención del alumno	0,691
Muestra interés por las inquietudes del alumno	0,663
Dialoga con los alumnos	0,662
Manifiesta entusiasmo por los contenidos	0,649
Fomenta la participación del alumno	0,643
Domina la docencia de la asignatura	0,605
Las actividades prácticas de la materia me han resultado útiles	0,603
Realiza una distribución adecuada del tiempo	0,558
En cada clase hace una presentación	0,515
Se ajusta con precisión al programa	0,479
Los materiales son de ayuda para preparar la materia	0,460
Emplea esquemas, gráficos, medios audiovisuales para apoyar explicaciones	0,426
El profesor cumple horario	0,375
La dificultad percibida de la asignatura es	0,194

c) El análisis factorial proporcionó resultados contradictorios en las dos fases del procedimiento. El análisis factorial realizado en el procedimiento piloto indicó una estructura bidimensional. Se podía considerar que el instrumento inicial media dos dimensiones de eficacia docente: la dimensión *Competencia docente* y la dimensión *Capacidad de interacción del profesor con el alumno*. El análisis factorial general realizado con la muestra total no detectó esta misma estructura bidimensional, detectándose un único factor general saturado por trece de las veinte variables contrastadas.

Autores como Villa (1993) afirman que lo deseable es que aparezcan diversos factores, dos al menos. Un único factor puede identificarse con calidad global, interpretándose en este caso que los alumnos no distinguen bien unos aspectos de otros y que evalúan sin matizar.

d) Los índices de fiabilidad que se obtienen cuando seguimos el procedimiento tradicional de análisis, pueden observarse en la Tabla 4.

TABLA 4: *Tabla resumen de los valores Alpha de Cronbach obtenidos*

Curso	Nº de casos	Nº de ítems	Alpha
97-98 1er semestre	3044	20	0,9375
97-98 2do semestre	2348	20	0,9538
98-99 1 er semestre	3874	20	0,9485
98-99 2 do semestre	2348	20	0,9458

Como puede comprobarse, los valores de fiabilidad obtenidos fueron estables en los cuatro periodos evaluados, y muy altos, con valores comprendidos entre un valor mínimo de 0,9357 en el primer semestre evaluado y un valor máximo de 0,9538.

La perspectiva de este trabajo es mostrar que estos resultados no proporcionan información suficiente respecto de la naturaleza del proceso de medición. En realidad en éste proceso hay dos componentes distintos. Uno el referido a la medida en que el instrumento refleja las opiniones de los alumnos y otro referido a la medida en que los alumnos son capaces de valorar una característica objetiva del profesor.

Metodológicamente esto se plasma en la necesidad de utilizar una estrategia distinta de la tradicional, en la que sea posible distinguir entre esos dos elementos. La herramienta metodológica adecuada viene facilitada por los modelos jerárquicos lineales.

Los modelos jerárquicos lineales pueden contemplarse como sistemas de ecuaciones de regresión jerárquicas y tienen como particularidad el hecho de que los coeficientes de regresión en un nivel, se convierte en variables dependientes en el siguiente. Para más información sobre los modelos jerárquicos lineales puede consultarse Bryk y Raudenbush (1992) y Bosker y Snijders (1999).

El modelo jerárquico que se describe a continuación pone en relación el factor *Competencia docente* y el factor *Capacidad de comunicación* del profesor con el alumno, con la eficacia docente.

Discusión de la naturaleza de las variables consideradas

Obtener la medición de la percepción del alumno sobre la eficacia docente de un profesor es relativamente sencillo. Pero si lo que se pretende es obtener una evaluación del profesor más objetiva, el procedimiento se complica. Se debe tener en cuenta que se trabaja con una definición operativa del constructo eficacia docente. Esta definición conceptualiza una serie de conductas que aparecen en el aula. Pero *crear* un constructo no es lo mismo que medirlo. Los constructos sólo pueden medirse indirectamente y la medida obtenida siempre tiene error. Es por lo tanto necesario cuantificar el grado en que los problemas de medida pueden influir en el procedimiento de evaluación que estamos utilizando.

Es necesario por tanto ampliar las fuentes de variación objeto de estudio e incluir las siguientes en el modelo:

a) la variación atribuida a la *incapacidad* del instrumento de captar con total exactitud la opinión del alumno.

b) la variación debida a la incapacidad del alumno de reflejar con total exactitud la verdadera característica del profesor.

c) la variación debida a la varianza verdadera entre profesores.

Teniendo en cuenta estos niveles de variabilidad se propone un modelo de tres niveles: nivel ítem, nivel alumno y nivel profesor, que pone en relación la competencia docente y la capacidad de interacción del profesor con la eficacia docente percibida por el alumno.

Descripción del modelo

En el Nivel ítem o primer nivel el modelo parte de la hipótesis de que la puntuación de un alumno en un ítem concreto (y_{ijk}) es una estimación insesgada de la verdadera opinión del alumno respecto del profesor y se definirá a través de una media que el alumno atribuye a un profesor en cada uno de los factores evaluados (B_{01jk} B_{02jk}) más un error de medida cometido en esa valoración (e_{1ijk}, e_{2ijk}).

Como el número de ítems es distinto para cada factor, existirán 2 términos de error asociados cada uno a la valoración en ese factor.

Ecuación en el Nivel 1 o Nivel ítem

$$Y_{ijk} = \sum_{p} \pi_{pijk} a_{pijk} + \sum_{p} e_{pijk}$$

$$Y_{ijk} = \beta_{01jk} a1_{ijk} + \beta_{02jk} a2_{ijk} + e1_{ijk} a1_{ijk} + e2_{ijk} a2_{ijk}$$

Donde

Y_{ijk} Es la puntuación en el ítem i del alumno j al profesor k

β_{01jk} Verdadera opinión sobre la competencia docente del profesor k del alumno j

β_{02jk} Verdadera opinión sobre la capacidad de interacción del profesor k del alumno j

e_{ijk} Error de medida del ítem respecto a la percepción del alumno

a_{ijk} Variable dummy que controla la pertenencia del ítem al factor. Si el ítem pertenece al factor 1 y estamos considerando competencia docente tomará el valor 1 en caso contrario tomará el valor 0.

En el Nivel alumno o segundo nivel el modelo parte de la hipótesis de que la media que un alumno da a un profesor en cada uno de los factores, es una estimación insesgada e incluye una media general que obtendría ese profesor más un termino de error. El error en este caso no es un error de medida del instrumento sino un error de percepción del alumno, que no puede cuantificar con total exactitud ese constructo.

Ecuación en el Nivel 2 o nivel ítem

$$\beta_{01jk} = \beta_{01k} + \mu_{01jk}$$

$$\beta_{02jk} = \beta_{02k} + \mu_{02jk}$$

Donde

β_{01jk} La verdadera opinión del alumno j sobre el profesor k en el factor competencia

β_{01k} Verdadero valor del profesor K en competencia docente

μ_{01jk} Error de percepción del alumno j sobre la competencia docente del profesor K

β_{02jk} La verdadera opinión del alumno j sobre el profesor k en el factor interacción

β_{02k} Verdadero valor del profesor K en capacidad de interacción

μ_{02jk} Error de percepción del alumno j sobre la competencia docente del profesor K

En el Nivel profesor o tercer nivel el modelo parte de la hipótesis de que la media de un profesor en cada uno de los factores depende de una media general en ese factor más un componente específico de ese profesor. Ese componente es lo que se diferencia ese profesor respecto a la media de todos los profesores.

Ecuación en el Nivel profesor o nivel 3

$$\beta_{01k} = \beta_{01} + \mu_{01k}$$

$$\beta_{02k} = \beta_{02} + \mu_{02k}$$

Donde

β_{01k} Verdadero valor del profesor K en competencia docente

β_{01} Media general en competencia docente

μ_{01k} Componente específico del profesor K

β_{02k} Verdadero valor del profesor K en capacidad de interacción

β_{02} Media general en interacción docente

μ_{02k} Componente específico del profesor k

Los parámetros que se obtienen del modelo son 2 interceptos y seis varianzas asociadas a los tres niveles de datos objeto de análisis:

$$Y_{ijk} = \beta_{oijk}a1_{ijk} + \beta_{1ijk}a2_{ijk}$$

$$\beta_{oijk} = \beta_0 + v_{0k} + \mu_{ojk} + \varepsilon_{oijk}$$

$$\beta_{1ijk} = \beta_1 + v_{1k} + \mu_{1jk} + \varepsilon_{1ijk}$$

$$\begin{bmatrix} V_{ok} \\ V_{1k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} \sigma_{v0}^2 & \\ 0 & \sigma^2_{v1} \end{bmatrix}$$

$$\begin{bmatrix} \mu_{ok} \\ \mu_{1k} \end{bmatrix} \sim N(0, \Omega_\mu) : \Omega_\mu = \begin{bmatrix} \sigma_{\mu 0}^2 & \\ 0 & \sigma^2_{\mu 1} \end{bmatrix}$$

$$\begin{bmatrix} \varepsilon_{ok} \\ \varepsilon_{1k} \end{bmatrix} \sim N(0, \Omega_\varepsilon) : \Omega_\phi = \begin{bmatrix} \sigma_{\varepsilon 0}^2 & \\ 0 & \sigma^2_{\varepsilon 1} \end{bmatrix}$$

La obtención de los índices de fiabilidad se realiza en dos etapas:

a) En una primera se calculan las medias en competencia e interacción docente, así como las varianzas atribuibles a cada uno de los niveles del modelo.

b) En segunda fase se calcula una medida de la fiabilidad de las medias obtenidas por cada profesor, a través de cocientes de varianzas entre los distintos niveles de datos como se explicará a continuación.

Resultados y discusión

La Tabla 5 muestra los datos del primer cuatrimestre del curso 97-98. Se trabajó con 159 unidades de nivel 3 es decir 159 profesores, 3045 alumnos o unidades de nivel 2 y 36540 ítems.

TABLA 5: *Interceptos del modelo que analiza la estructura del cuestionario. Curso 97-98. Primer cuatrimestre*

Año Académico	Semestre	Factor	Intercepto	Error
97-98	1	A1	5,315	0,053
		A2	4,987	0,061

La media obtenida por todos los profesores en el factor competencia docente es de 5,315 mientras que la obtenida en el factor interacción con el alumno es un poco más baja de 4,987, siendo ambas significativas.

Los intervalos de confianza de estas puntuaciones son:

$$5{,}315 \pm 1{,}96 * 0{,}053 = (5{,}41,\ 5{,}21); \quad 4{,}978 \pm 1{,}96 * 0{,}061 = (5{,}09\ ,\ 4{,}8)$$

La Tabla 6 muestra los valores de las varianzas en cada uno de los tres niveles. Estos son significativamente mayores que 0, lo cual nos indica que existe una variación significativa entre los 3045 alumnos y entre los 159 profesores.

TABLA 6: *Varianzas del modelo que analiza la estructura del cuestionario. Curso 97-98. Primer cuatrimestre*

Año Académico	Semestre	Factor	Nivel	Varianza	Error
97-98	1	A1/A1	3	0,3227	0,048
		A2/A2	3	0,4223	0,064
		A1/A1	2	1,330	0,039
		A2/A2	2	1,872	0,053
		A1/A1	1	1,184	0,011
		A2/A2	1	0,941	0,014

Las estimaciones indican que la mayor variación en los resultados se produce en el nivel alumno especialmente al valorar la capacidad de interacción del profesor con los alumnos.

Los datos de ese mismo curso académico pero en el segundo semestre aparecen en la Tabla 7. Se evalúan 170 profesores, 2348 alumnos contando con 28176 ítems.

TABLA 7: *Interceptos del modelo que analiza la estructura del cuestionario. Curso 97-98. Segundo cuatrimestre.*

Año Académico	Semestre	Factor	Intercepto	Error
97-98	2	A1	5,314	0,063
		A2	4,950	0,074

En el segundo semestre también se encuentran medias significativas en el factor Competencia Docente (*5,314*), y en el factor interacción (*4,950*).

Los intervalos de confianza son:

$$5{,}13 \pm 1{,}96 * 0{,}063\ (\ 5{,}15,\ 5{,}00); \ 4{,}95 \pm 1{,}96 * 0{,}074\ (5{,}09,\ 4{,}8)$$

La varianza obtenida vuelve a ser significativa en los tres niveles, existiendo diferencias significativas entre los profesores, la valoración de los alumnos y entre los ítems. En el nivel profesor aumenta la varianza, en el nivel alumno disminuye ligeramente, y se mantiene entre los ítems como puede observarse en la Tabla 8

TABLA 8: *Varianzas del modelo que analiza la estructura del cuestionario. Curso 97-98. Segundo cuatrimestre*

Año Académico	Semestre	Factor	Nivel	Varianza	Error
97-98	2	A1/A1	3	0,548	0,075
		A2/A2	3	0,753	0,103
		A1/A1	2	1,196	0,041
		A2/A2	2	1,598	0,055
		A1/A1	1	1,191	0,013
		A2/A2	1	0,917	0,015

En el curso académico 98-99 se cuenta con 235 profesores, 3884 alumnos y 46488 ítems a comprobar. La Tabla 9 muestra sus resultados

TABLA 9: *Interceptos del modelo que analiza la estructura del cuestionario. Curso 98-99. Primer cuatrimestre*

Año Académico	Semestre	Factor	Intercepto	Error
98-99	1	A1	5,221	0,057
		A2	4,868	0,066

En el curso académico 98-99 las medias en los factores competencia docente (5,221) e interacción con el alumno (4,866) siguen siendo significativas. Los intervalos de confianza son:

$$5{,}221 \pm 1{,}96*0{,}057\ (5{,}33\ \ 5{,}10);\ 4{,}868 \pm 1{,}96*0{,}066\ (4{,}99\ \ 4{,}76)$$

Las varianzas asociadas a las medias son significativas en los tres niveles como puede apreciarse en la Tabla 10

TABLA 10: *Varianzas del modelo que analiza la estructura del cuestionario. Curso 98-99. Primer cuatrimestre.*

Año Académico	Semestre	Factor	Nivel	Varianza	Error
98-99	1	A1/A1	3	0,689	0,072
		A2/A2	3	0,901	0,096
		A1/A1	2	0,997	0,027
		A2/A2	2	1,442	0,040
		A1/A1	1	1,191	0,010
		A2/A2	1	0,965	0,013

La Tabla 11 muestra los resultados en el último semestre. Las medias obtenidas de 2348 alumnos y 174 profesores fueron más altas tanto en competencia docente (5,406) como en interacción (5,269). El intervalo de confianza es:

$$5{,}406 \pm 1{,}96*0{,}06\ (5{,}5\ \ 5{,}2);\ \ 5{,}269 \pm 1{,}96*0{,}08\ (5{,}42\ \ 5{,}11)$$

TABLA 11: *Interceptos del modelo que analiza la estructura del cuestionario. Curso 98-99. Segundo cuatrimestre*

Año Académico	Semestre	Factor	Intercepto	Error
98-99	2	A1	5,406	0,06895
		A2	5,269	0,08816

Se mantiene, como puede verse en la Tabla 12 la significatividad en las varianzas en los 3 niveles:

TABLA 12: *Varianzas del modelo que analiza la estructura del cuestionario. Curso 98-99. Segundo cuatrimestre.*

A. Académico	Semestre	Factor	Nivel	Varianza		Error	
98-99	2		A1/A1	3	0,738		0,088
			A2/A2	3	1,224		0,145
			A1/A1	2	0,724		0,026
			A2/A2	2	1,045		0,037
			A1/A1	1	1,095		0,012
			A2/A2	1	0,747		0,013

Las Tablas 13 y 14 muestran el resumen de los resultados obtenidos en las medias y en las varianzas y nos permiten afirmar que:

- Las medias de los 2 factores son significativas en los 4 semestres evaluados siendo la competencia docente más valorada por los alumnos *(5,315, 5,314, 5.221, 5,406)* que la capacidad de interacción de sus profesores *(4,98, 4,95, 4,86, 5,26).*

- Las puntuaciones se han mantenido estables en el tiempo. Los parámetros fijos resultan todos significativos, ya que si dividimos el valor del estimador por su error típico el cociente resultante es siempre superior a 2, que es el valor utilizado para el nivel de significación del 0,05. Por lo tanto podemos afirmar que tanto la competencia docente como la capacidad de interacción con el alumno son predictores de la eficacia docente percibida en los tres niveles.

- Las varianzas asociadas a los tres niveles son significativas. Este modelo explica gran parte de la varianza entre las medias de los profesores, de las varianzas de las medias de los alumnos y entre las medias de los ítems.

- Los profesores difieren entre sí en competencia e interacción docente, especialmente en ésta última (*1,244*).

- Los alumnos difieren entre sí al valorar la competencia de la interacción docente y los ítems utilizados también discriminan bien estos conceptos, especialmente la competencia docente (*1,095*).

TABLA 13: *Tabla resumen. Interceptos obtenidos del modelo que analiza la estructura del cuestionario*

	Competencia Docente	Interacción
97-98 1c	5,315	4,987
97-98 2c	5,314	4,950
98-99 1c	5,221	4,868
98-99 2c	5,406	5,269

TABLA 14: *Tabla resumen. Varianzas obtenidas en el modelo que analiza la estructura del cuestionario*

		97-98 1c	97-98 2c	98-99 1c	98-99 2c
Profesor	Competencia	0,323	0,547	0,689	0,783
	Interacción	0,422	0,753	0,901	1,224
Alumnos	Competencia	1,330	1,196	0,997	0,724
	Interacción	1,782	1,598	1,442	1,045
Items	Competencia	1,184	1,191	1,191	1,095
	Interacción	0,941	0,917	0,965	0,747

Obtención de la Fiabilidad de las medias

Una vez conocidas las medias generales en competencia e interacción docente y el reparto de varianza que es debido a la variación entre alumnos y entre profesores pretendíamos a través del análisis de los errores conocer una medida de la fiabilidad de las medias obtenidas por los profesores con este modelo.

Nuestro objetivo es poder confirmar que las diferencias obtenidas se deben fundamentalmente a que los profesores son distintos y así son percibidos por los alumnos y no a que los alumnos son distintos en forma de evaluar.

Los coeficientes de fiabilidad se han calculado con cocientes de varianza entre los diversos niveles y su objetivo es explicar los porcentajes de varianza de las puntuaciones atribuibles a cada nivel. Nos interesa conocer:

a) La Fiabilidad del proceso al medir al profesor:

$$\frac{Varniv3}{Varniv3+Varniv2+Varniv1}$$

b) La Fiabilidad del alumno al medir al profesor:

$$\frac{Varniv3}{Varniv3+Varniv2}$$

c) La Fiabilidad del instrumento al medir al alumno:

$$\frac{Varniv2}{Varniv1+Varniv2}$$

d) El índice de fiabilidad comparable al Alpha de Cronbach:

$$\frac{Varniv2+Varniv3}{Varniv1+Varniv2+Varniv3}$$

La comparación de las estimaciones de varianza entre los tres niveles pretende explicar la distribución de los componentes de varianza que se produce al introducir varios niveles.

La Tabla 15 resume las varianzas estimadas entre profesores, alumnos e ítems en *competencia docente*:

TABLA 15: *Varianzas obtenidas en Interacción docente en el modelo que analiza la estructura del cuestionario*

	97-1c	97-2c	98-1c	98-2c
Profesor – Instrumento	0,114	0,186	0,239	0,301
Alumno –Profesor	0,195	0,314	0,409	0,520
Instrumento –Alumno	0,529	0,501	0,456	0,398
Comparable al Alpha	0,582	0,594	0,586	0,579

Si observamos la primera fila de la tabla que muestra los niveles de varianza debidos a la variabilidad entre profesores comprobamos que, como máximo un 30,1% de la varianza (y sólo en uno de los períodos evaluados), es debida exclusivamente a las diferencias entre profesores.

La segunda fila de la tabla muestra la varianza debida a las diferencias entre los alumnos a la hora de evaluar. En este caso es atribuible como máximo un 52% de la varianza obteniéndose resultados poco estables. La tercera fila muestra la varianza que es atribuible al instrumento. El valor máximo es de un 52,9% observándose una mayor estabilidad en los periodos evaluados. Por último la cuarta fila muestra el índice comparable al Alpha de Conbrach, en el cual se observa una disminución de la fiabilidad.

Respecto a la Interacción docente obtuvimos los resultados que aparecen en la Tabla 16.

TABLA 16: *Varianzas obtenidas en Interacción docente en el modelo que analiza la estructura del cuestionario*

	97-1c	97-2c	98-1c	98-2c
Profesor-Instrumento	0,140	0,230	0,272	0,406
Alumno-Profesor	0,199	0,320	0,385	0,539
Instrumento-Alumno	0,654	0,635	0,599	0,583
Comparable al Alpha	0,703	0,719	0,708	0,752

La primera fila de la tabla muestra los niveles de varianza obtenidos en los cuatro periodos evaluados debidos a la variabilidad entre profesores en interacción docente. Comprobamos que como máximo un 40,6% de la varianza (y sólo en uno de los periodos evaluados) es debida exclusivamente a la diferencia entre profesores.

La segunda fila de la tabla muestra la varianza debida a la diferencia entre los alumnos a la hora de evaluar. En este caso es atribuible como máximo un 53,9% de la varianza, obteniéndose resultados poco estables.

La tercera fila muestra la varianza que es atribuible al instrumento. El va-

lor máximo en interacción docente es de un 65,4 %, obteniéndose valores de una mayor estabilidad. El índice comparable al Alpha de Cronbach que aparece en la última fila también indica una disminución de la fiabilidad.

Es decir, de las fuentes de variación utilizadas la información más poderosa en el conocimiento de la eficacia docente es la ofrecida por el instrumento a los alumnos con una varianza de 65,4% en interacción docente y un 52,9% en competencia docente.

Es más dudosa la capacidad de discriminación de la eficacia docente de los alumnos sobre sus profesores (con varianzas máximas del 52,0% y 53,9%) y del instrumento sobre los profesores evaluados (30,1% y 40,6%).

Estos valores contrastan fuertemente con los valores de fiabilidad obtenidos en el apartado anterior y en otras investigaciones que cifran la fiabilidad de este tipo de instrumentos en valores superiores a 0,90.

La ampliamente aceptada capacidad de los alumnos como fuente de información fiable de sus profesores (Marsh (1987), Aparicio et al. (1982) y Tejedor (1993) entre otros), es cuestionable a la vista de los resultados obtenidos.

Conclusiones

Los valores de *fiabilidad* del instrumento calculados a través del A*lpha* de Cronbach en el análisis exploratorio parecían indicar que estábamos trabajando con un instrumento cuya consistencia interna era muy elevada, con valores superiores a *0,90* en todos los periodos evaluados.

Las estimaciones obtenidas bajo los supuestos de los modelos jerárquicos lineales, considerablemente inferiores, no permiten ser tan optimistas respecto a la fiabilidad del instrumento.

Esta disminución ha sido debida a dos factores: en primer lugar a la descomposición del coeficiente de fiabilidad en función de las diversas jerarquías y agrupamientos en los datos, y en segundo lugar, a su obtención a través de cocientes de varianza entre los niveles de datos.

El objetivo del cálculo de los cocientes de varianza a nivel ítem, alumno y profesor era determinar si los ítems del cuestionario diseñado sirven para detectar profesores eficaces, y si los alumnos a través de ellos son capaces de valorar estas cualidades en determinados profesores.

Los porcentajes de varianza obtenidos permiten afirmar que el instrumento está técnicamente bien diseñado y que es una fuente de información *relativamente* fiable (con valores en torno al 65%). Pero también permiten dudar sobre su capacidad para discriminar docente eficaces y dudar sobre la capacidad de discriminación del alumno sobre la capacidad de interacción docente, y especialmente, sobre la competencia docente.

No podemos olvidar que nuestro mo-

delo estaba diseñado para estimar si los errores de medida inherentes a cualquier proceso de medición se veían incrementados por los errores debidos a la medición indirecta de un constructo operativizado a través de conductas difíciles de evaluar.

Estos errores, originados por la dificultad de cuantificar numéricamente una característica no basada en cualidades físicas, sino en percepciones de carácter muy subjetivo, sí han tenido peso a la hora de estimar la fiabilidad del alumno como evaluador de la eficacia docente. Los resultados muestran una reducción considerable de todos los índices de fiabilidad.

Ante los resultados anteriormente descritos se puede afirmar que es necesario la optimización y mejora de estos procedimientos de evaluación, ya que no se puede confirmar que las diferencias obtenidas sean debidas a que los profesores son distintos y así son percibidos por los alumnos.

No obstante, dada la naturaleza de los datos y la limitación en el número de variables utilizadas en los distintos niveles (sólo se han podido utilizar dos predictores en el primer nivel) es deseable añadir algunas variables que pueden tener influencia cuando se utiliza a los alumnos como fuente de información de la eficacia docente.

Deben considerarse variables como el tamaño de la clase, el tipo de asignatura, el sexo de los alumnos, el tipo de instrucciones en la aplicación del instrumento y la motivación de los alumnos entre otras.

Las limitaciones encontradas en este estudio no tratan de descartar el uso de las encuestas de evaluación en el conocimiento de la eficacia docente. Se trata por el contrario, de mejorar el tipo de hipótesis y relaciones a tener en cuenta en la consideración del alumno como fuente de información sobre la calidad docente.

Se comprueba que la evaluación de la docencia universitaria a través del alumno no es un sistema perfecto, pero creemos que es uno de los mejores sistemas. El alumno es el único capaz de juzgar si la docencia recibida le ha ayudado a aprender y este hecho fundamental convierte su percepción de la eficacia docente, a pesar de las limitaciones, en necesaria y útil.

A pesar de esto, este trabajo pretende demostrar que esa no puede ser la única fuente de información en la evaluación de la competencia docente por los motivos aquí desarrollados.

Dirección de los autores: Isabel Pascual Gómez, Universidad Antonio de Nebrija, Campus de la Berzosa 28240, Hoyo de Manzanares (Madrid), ipascual@nebrija.es.

Fecha de recepción de la versión definitiva de este artículo: 15.IX. 2004

Bibliografía

APARICIO, J. J., SAN MARTIN, R. y TEJEDOR, F. J. (1982) *La enseñanza universitaria vista por los alumnos: Un estudio para la evaluación de los profesores en la enseñanza superior,* Madrid, I. C. E. Universidad Autónoma.

APODAKA, P. y RODRÍGUEZ, M. (1999) La Opinión de los Alumnos en la Evaluación de la Calidad Docente: Posibilidades, Limitaciones, y Estructura dimensional de sus Indicadores, *Indicadores en la Universidad: información y decisione*s, pp. 311-328, Madrid, Consejo de Universidades.

BALLANTYNE, R., BORTHWICK, J. y PACKER, J. (2000) Beyond Student Evaluation of Teaching: Identifying and Addressing Academic Staff Development Needs, *Assesment & Evaluation in Higher Education,* 25:3, pp. 221-236.

BOSKER, R. y SNIJDERS, T. (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (London, Sage).

BRYK, S. y RAUDENBUSH, W. (1992) *Hierarchical Linear Models* (Sage, California).

BRUCE, A. J. (1985) A Comparison of Three Teaching Evaluation Instruments. Convention of the Southwestern Psychological Association. April.

COHEN, P. A. (1980) Effectiveness of Student-Rating Feedback for Improving College Instruction: A Meta-analysis of Findings, *Research in Higher Education*, 13, pp. 312-341.

CRUSE, D**.** (1987) Student Evaluations and the University Professor, *Higher Education*, 15:6, pp. 723-737.

FELDMAN K. A. (1996) Identifying Exemplary Teaching: Using Data from Course and Teacher Evaluation, *New Directions for Teaching and Learning,* 65, pp. 41-50.

FERNÁNDEZ, J. y MATEO, M. (1992) Student Evaluation of University Teaching Quality: Analysis of Questionnaire for a Sample of University Students in Spain, *Educational and Psychological Measurement,* 52, pp. 675-684.

GOLDSTEIN, H. (1986) Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares, *Biometrika,* 73, pp. 43-56.

GOLDSTEIN, H**.** (1995) *Multilevel Statistical Models* (London, Arnold).

HASKELL, R. E**.** (1997) Academic Freedom, Tenure, and Student Evaluation of Faculty: Galloping Polls in the 21st Century. [en línea] *Education Policy Analysis Archives***,** 5:6, February, http://www.bus.lsu.edu/accounting/faculty/lcrumbley/educpoly.htm, [Consulta: 5 marzo 2001].

HOX, J. J. (1995) *Multilevel Modeling: When and Why*. [En línea] (Amsterdam. TT-Publikaties) September, http://www.fss.uu.nl/ms/jh/publist/whenwhy.pdf, [Consulta Julio 2001].

MARSH, H. W. (1987) Students Evaluation of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research, *International Journal of Educational Research*, 11:3, pp. 253-388.

OCAÑA-RIOLA, R. (1999) Métodos Estadísticos en la Gestión y Política Sanitaria. [En línea] *Revista de Administración Sanitaria*, 3:9, http://www.dinarte.es/ras/ras09/prin09.htm, [Consulta Enero 2002].

RAUDEBUSH, S., ROWAN, B. y CHEONG, F. Y**.** (1991) A Multilevel, Multivariate Model for School Climate with Estimation Via the EM Algorithm and Aplicattion to US high School Data, *Journal of Educational Statistics*, 16, pp. 295-330.

RAUDENBUSH, S. W. (1988) Educational Applications of Hierarquical Linear Models: A Review, *Journal of Educational Statistics*, 13:20, pp. 85-116.

SPROULE, R**.** (2000) Student Evaluation of Teaching: A Methodological Critique of Conventional Practices [En línea] 2000, *Education Policy Archives*, 8 (50) pp. 1-21, November, http://epaa.asu.edu/epaa/v8n50.html, [Consulta Marzo 2001]

TEJEDOR F. J. (1993) Experiencias Españolas de Evaluación de la Enseñanza Universitaria y Nuevas Perspectivas**,** *III Jornadas Nacionales de Didáctica Universitaria*, pp. 85-109, Universidad de las Palmas de Gran Canaria.

VILLA, A. y MORALES, P. (1993*) La evaluación del profesor. Una visión de los principales problemas y enfoques en diversos contextos* (Vitoria, Servicio Central de Publicaciones del Gobierno Vasco).

VILLA SÁNCHEZ, A. y VILLARDON GALLEGO, L. (1995) Validez de la Evaluación Docente Universitaria a Través de los Alumnos, *Planificación, Evaluación y Financiación de los sistemas educativos***,** pp. 267-280, Jornadas AEDE**.**

WEISSSBERG, R. (1993) Managing Good Teaching, *Perspectives on Political Sciences*, 22:1, pp. 21-28.

Resumen:
El problema de la fiabilidad en la evaluación de la eficacia docente en la universidad: una alternativa metodológica

Este artículo muestra que cuando los índices de fiabilidad de un cuestionario de evaluación docente en la universidad se descomponen y se separan los componentes personales del alumno y los componentes contextuales se produce una considerable reducción de la fiabilidad. Los resultados sugieren la necesidad de considerar otras alternativas metodológicas respecto de la fiabilidad calculada por los métodos tradicionales, e interpretar los resultados obtenidos de las encuestas con cautela, especialmente si van a ser utilizadas por los responsables académicos con fines sumativos.

Descriptores: Evaluación docente en la Universidad, fiabilidad de los cuestionarios a alumnos, eficacia de los profesores.

Summary:
The problem of reliability when evaluating teaching efficiency: a methodological alternative

This article deals with the following issue: When reliability indices of teaching evaluation at university level are broken down into students' personal components and contextual components, reduction of reliability is expected. Results suggest that it is necessary to consider a methodological alternative to reliability calculated with traditional methods, and to interpret cautiously results obtained from questionnaires, especially if they are going to be used for summative assesment by members of the academic authorities.

Key Words: Hierarchical linear models, reliability indices, teaching effectiveness.