Sobre la evaluación del profesorado universitario. Una réplica a Muñiz y Fonseca-Pedrero

Andrés González Gómez
andreito@ugr.es
Departamento de Psicología Social y Metodología de las Ciencias del Comportamiento
Facultad de Psicología
Universidad de Granada

RESUMEN: Se analiza el proceso de evaluación universitaria tomando como guía el trabajo de Muñiz y Fonseca-Pedrero (2008). Se plantea que la mayoría de procesos de evaluación universitaria presentan una confusión en la aproximación al escalamiento utilizada, mezclándose técnicas de elaboración de cuestionarios, medidas centradas en los estímulos y metodología de encuestas. Se destaca la indefinición del papel que juegan los distintos agentes implicados en la evaluación, resultando complicado determinar quién (profesores o alumnos) es el evaluador y quién el evaluado.

PALABRAS CLAVE: Evaluación universitaria. Construcción de test. Escalamiento.

On the evaluation of teachers college. A reply to Muñiz and Fonseca-Pedrero

ABSTRACT: The proccess of university evaluation is analized following the work of Muñiz and Fonseca-Pedrero (2008). The aim of the paper is to point out the methodological confusion in the majority of evaluation processes at universities, as a result of a misleading scaling approach as well as the mixture of several techniques of questionnaire design and survey methodology. The paper highlights the lack of definition with regard to the role of the people involved in the evaluation, which makes it difficult to distinguish who is being evaluated and who is de evaluator.

KEY WORDS: University evaluation. Test construcction. Scaling.

Fecha de recepción: 18/06/2009 · Fecha de aceptación: 18/09/2009
Dirección de contacto:
Andrés González Gómez
Departamento de Psicología Social y Metodología de las Ciencias del Comportamiento
Facultad de Psicología
Universidad de Granada
18071 Granada

1. INTRODUCCIÓN

En un reciente artículo en esta misma revista, Muñiz y Fonseca-Pedrero (2008) plantean la secuencia de acciones que deben acometerse en el proceso de construcción de un instrumento para la evaluación universitaria. Como paso previo, plantean los requisitos y el modelo general que debe seguir cualquier proceso de evaluación. Estoy totalmente de acuerdo con los autores en que pareciera existir un fantasma de la evaluación que recorre todas las universidades españolas. Ahora bien, entre las propiedades de ese fantasma no estaría la de cambiar de aspecto, dada la elevada homogeneidad –al menos en apariencia– de las evaluaciones que se realizan en todas estas instituciones (Muñoz Cantero, Ríos de Deus y Abalde, 2002). Lamentablemente, la coincidencia o estabilidad en los procedimientos utilizados no garantiza su validez. El trabajo de Muñiz y Fonseca-Pedrero creo que es un buen reflejo del estado de la cuestión, o de cuál es la filosofía subyacente en las evaluaciones realizadas en las universidades españolas. Pero, en mi opinión, esta

visión ortodoxa y tradicional del proceso de elaboración de instrumentos de medida no por ello se adapta correctamente a las situaciones de evaluación universitaria.

En este artículo trataré de señalar cuáles son los problemas que esta aplicación directa conlleva. Para ello iré analizando los pasos que Muñiz y Fonseca-Pedrero plantean en su trabajo. No pretendo con ello realizar una crítica específica a la propuesta de estos autores. Al contrario, creo que su aportación resume perfectamente la forma de proceder –en el mejor de los casos– que siguen los responsables de la evaluación de las distintas universidades. Por tanto mis comentarios van dirigidos, no sólo a Muñiz y Fonseca-Pedrero, si no todos los interesados en los procesos de evaluación que, en la práctica, parecen suscribir sus tesis.

2. EL PROCESO GENERAL DE EVALUACION

La primera parte del trabajo de Muñiz y Fonseca-Pedrero resume siete cuestiones clave que se deben incluir en cualquier modelo general de evaluación. Estas son: qué se evalúa; cuáles son las partes legítimamente implicadas en la evaluación; quién evalúa; cómo se evalúa; qué metodología utilizar; qué *feedback* se ofrece a las partes implicadas; planes de mejora generados por la evaluación y opinión de las partes implicadas sobre la evaluación. Pocas pegas se pueden poner a la pertinencia de estas cuestiones. Es más, un análisis de lo que viene aconteciendo en nuestras universidades permite anticipar cuáles son las respuestas habituales a estas preguntas. De este modo y al menos en parte, podemos decir que en la mayoría de los casos el qué son los profesores; el quién los alumnos y el cómo mediante un cuestionario.

Pero, si ninguna de las respuestas individuales dadas a estas preguntas es equivocada, su confluencia deviene en problemática. En mi opinión este factor ha pasado habitualmente desapercibido y es una fuente de incongruencias que trataré de mostrar a continuación. Centraré mis comentarios en la segunda parte del trabajo que es la que constituye el núcleo principal del mismo. En relación a la primera parte tan sólo me gustaría señalar que la afirmación de los autores de que: "los instrumentos que se utilicen para la evaluación han de ser objetivos, claros, comprensibles por las partes, preferiblemente cuantitativos, fiables y válidos" (p. 15), puede no ser tan conveniente como aparenta. Desde luego, la validez y la fiabilidad no son características intrínsecas de los instrumentos (American Educational Research Association, American Psychological Association y National Council on Measurement in Education, 1999). Y este punto es más importante si se tiene en cuenta que son varios los agentes que pueden hacer uso de los resultados de la evaluación.

3. LA ELABORACION DE INSTRUMENTOS DE EVALUACION UNIVERSITARIA

La segunda parte es la dedicada específicamente a la elaboración de instrumentos para la evaluación universitaria. Comienza señalando que: "En el contexto de la evaluación universitaria, entendemos por instrumento de medida un procedimiento estandarizado que permite obtener un conocimiento objetivo de una persona, producto, sistema o institución". Es decir, a priori, no se centra exclusivamente en la evaluación de personas, sino que plantea como posible objeto de medida del instrumento de evaluación también a los productos, los sistemas y las instituciones. Esta distinción es altamente relevante dado que puede generar el decantarse por una aproximación al escalamiento centrada en las respuestas o en los estímulos, en lugar de la clásica aproximación centrada en las personas (Torgenson, 1958).

La elección de una u otra aproximación general al escalamiento conllevará la selección de distintas técnicas. Así, si bien la aproximación centrada en las personas suele implicar la elaboración de instrumentos de medida de variables psicológicas (tal cual podría ser la "capacidad docente de los profesores") una aproximación basada, por ejemplo, en los estímulos implicará la elección de otra técnica de recogida de datos, como bien podría ser el caso de intentar ordenar una serie de instituciones (facultades, departamentos, etc.) en un presunto continuo (p.e. calidad docente). En el primer caso, cuando el interés de la medida es, digamos, un atributo que pueden mostrar en distinto grado los distintos profesores, una forma adecuada de medirlo sería mediante un test psicométrico. Como veremos más adelante, una de las características básicas de los test psicométricos es que consisten en obtener una muestra del comportamiento de las personas. Por el contrario, ni las instituciones, ni los sistemas ni los

productos tienen la posibilidad de mostrar esos comportamientos a partir de los cuales inferir su nivel en un constructo psicológico de interés. Con lo que se hace evidente la obligatoriedad de proceder de forma distinta según estemos interesados en una u otra aproximación a la medida.

Si bien hasta el momento Muñiz y Fonseca-Pedrero no se decantan en exclusiva por una u otra aproximación, la especificación de los pasos para elaborar un instrumento de evaluación, y las secciones siguientes, corresponden fielmente con los pasos habituales para la elaboración de test psicológicos, olvidándose de hecho de las otras posibilidades planteadas.

El resto de los apartados del trabajo de Muñiz y Fonseca-Pedrero es un recorrido ortodoxo de los que habitualmente se recomiendan en el proceso de construcción de test. En mi opinión, poco se puede criticar a esa serie de recomendaciones. El problema surge cuando, en el mejor de los casos, se aplican éstas en un proceso de evaluación sin una reflexión previa sobre su pertinencia. No me es posible, ni pretendo, juzgar las intenciones de los responsables de la evaluación de la docencia en las universidades españolas. Cabe esperar, además, suficiente variabilidad como para invalidar cualquier juicio colectivo. Sin embargo, en la mayoría de los casos, si no en todos, se puede apreciar una serie de regularidades que merece ser comentada. Esta regularidad comienza, como no, por el interés en medir la actuación docente del profesorado y continúa con la utilización de cuestionarios que son administrados a los alumnos con el objetivo de satisfacer el interés anterior. Ahora bien, esta forma de proceder presenta una serie de problemas que intentaremos aclarar a continuación.

El primer paso del proceso de construcción de instrumentos de medida es el que Muñiz y Fonseca-Pedrero denominan *Marco General del Instrumento de Medida*. Según estos autores, en este apartado adquiere especial importancia la determinación de "la población objeto de medición (alumnos, profesores, departamentos, etc.)" (Muñiz y Fonseca-Pedrero, 2008, p 17). Este primer aspecto, que podría parecer fácil de determinar pasa ser algo confuso e indeterminado en la mayoría de evaluaciones universitarias. En la evaluación docente en distintas universidades podemos encontrar distintas etiquetas que parecen jugar papel de constructo objeto de medición. Ahora bien, el nombre otorgado a un constructo no es más que una etiqueta resumen (Crocker y Algina, 1986). Si bien es plenamente posible, y así sucede en la práctica, que estas etiquetas correspondan a distintas concepciones y, por tanto, recojan distintos comportamientos, todas ellas comparten su concepción como constructos psicológicos. De otra forma no tendría sentido seguir los pasos para la elaboración de test en la medida de esas variables. En otras palabras, sólo si se acepta la naturaleza como variable psicológica del objeto de medida (la "calidad docente" del profesorado) tiene sentido proceder a su medida mediante la elaboración de un test.

La situación planteada es conceptualmente idéntica a –pongamos por caso– el interés en analizar el nivel de ansiedad o depresión o la capacidad de razonamiento lógico de una serie de profesores. Cada uno de estos constructos o variables psicológicas se manifestará de forma diferencial a través de la conducta de los profesores, y la mejor manera de medirlo será, obviamente, obteniendo una muestra de dicha conducta.

Frente a esto tenemos otra posibilidad que, estando más o menos relacionada, es claramente distinta. Me refiero a una situación donde el objetivo es conocer algo de los alumnos y no de los profesores. Ese algo puede ser su opinión sobre los planes de estudios, sobre los servicios de la universidad o sobre cualquier otro tema; el grado en que utilizan distintos servicios; su nivel de satisfacción con algún elemento, etc. En algunos casos, el objeto de medida será también un constructo psicológico y será necesaria de nuevo la elaboración de un test. Pero en este caso, las conductas de interés, las que manifiestan el constructo, serán conductas mostradas por los alumnos y no por los profesores.

Ambas situaciones son posibles –y deseables– en el contexto de la evaluación universitaria. Pero ambas situaciones deben estar claramente diferenciadas. Es decir, una cosa será el interés en constructos psicológicos de los profesores y otra el interés en constructos psicológicos de los alumnos. Esto que podría parecer una obviedad no lo es tanto cuando el esquema habitual consiste en asignar puntuaciones a los profesores mediante la administración de cuestionarios a los alumnos.

La confusión se hace patente, por ejemplo, en la Universidad de Granada. Recientemente se ha publicado un libro titulado "Opinión del alumnado sobre la actuación docente del profesorado de la Universidad de Granada. Resultados 2004-2007" (Defior y otros, 2007). El título comienza "opinión

del alumnado", lo que lleva a pensar que son los alumnos el objeto de medida. Sin embargo estas opiniones se refieren a la actuación docente de los profesores, lo que empieza a generar confusión. Esta mezcla se agita del todo cuando los que obtienen la puntuación en el cuestionario utilizado son en un primer nivel los profesores. En siguientes niveles también obtienen puntuación las áreas, departamentos y titulaciones.

En definitiva, además de otras consideraciones como el contexto de aplicación, formato, utilización futura de las puntuaciones, etc. el primer e ineludible paso en proceso de construcción de test, si este fuera el caso, es la determinación precisa de la población objeto de estudio. En adelante, supondré que la misma son los profesores universitarios y trataré de especificar las particularidades que esta opción acarrea. Si la elección fuese otra, por ejemplo los alumnos, las particularidades serían evidentemente distintas.

El siguiente paso consiste en la *Definición de la Variable Medida*. Como ya hemos apuntado antes, en el caso de los profesores, las variables de interés se pueden considerar constructos psicológicos y, como tales, es necesario definirlos desde distintas perspectivas. Por un lado será necesario establecer cuáles serán los comportamientos (de los profesores) que ponen de manifiesto la variable. Por otro lado será necesario anticipar las relaciones entre el constructo medido (en los profesores) y toda otra serie de variables. La especificación de estas variables constituye la teoría sobre el constructo que tiene el autor del test. En el caso que nos ocupa habría que especificar, por ejemplo, si cabe esperar una relación (y de qué tipo) entre las puntuaciones en el test y la categoría y antigüedad de los profesores. Es decir, si se espera, por ejemplo, que los profesores obtengan peores puntuaciones cuanto mayor es su categoría profesional; si se espera obtener diferencias en función de la licenciatura o el área de conocimiento y cuáles serán esas diferencias, etc. Toda esta serie de relaciones constituyen, como adecuadamente señalan Muñiz y Fonseca-Pedrero, la definición sintáctica del constructo.

Muñiz y Fonseca-Pedrero sintetizan adecuadamente la clave de la definición operacional de un constructo psicológico cuando señalan que es necesario "identificar con claridad los **comportamientos** más representativos de la variable de medición" (p. 19, la negrita es nuestra). Pero, ¿a quién se refieren esos comportamientos? Muñoz Cantero, Ríos de Deus y Abalde (2002) realizan un exhaustivo repaso de los diferentes cuestionarios utilizados en la evaluación del profesorado en las universidades españolas. Los autores recogen toda la diversidad de ítems utilizados, y entre ellos podemos encontrar ejemplos como:

- La asistencia a clase es una ayuda importante para la comprensión de la asignatura. (7)
- Los materiales de estudio (textos, apuntes, etc.) son adecuados. (10)
- En esta asignatura tenemos claro lo que se nos va a exigir. (9)
- En general, me siento satisfecho/a asistiendo a sus clases. (9)
- En sus explicaciones se ajusta bien al nivel de conocimiento de los estudiantes. (7)

Se hace patente la diversidad de aspectos (no sólo comportamientos de los profesores) que suelen recoger estos cuestionarios. Los números entre paréntesis indican el número de universidades que utilizan cada ítem en sus evaluaciones sobre un total de 17 analizadas en el trabajo de Muñoz Cantero, Ríos de Deus y Abalde, (2002). Esta breve muestra refleja claramente la confusión habitual en estas evaluaciones.

El tercer paso del proceso es el denominado *Especificaciones del Instrumento de Medida*, por Muñiz y Fonseca- Pedrero. En él, los autores plantean una serie de recomendaciones de carácter general que serían de aplicación en cualquier proceso de elaboración de cuestionarios. Sea este destinado a la evaluación del profesorado universitario o no. Por tanto, nada de lo que se comenta en este apartado puede ser catalogado como erróneo. No obstante, la particularidad de los cuestionarios a la que nos estamos refiriendo aconsejaría especificar cómo se concretan estas generalidades en el terreno de la evaluación del profesorado. Así, no parece a priori que el nivel cognitivo de los participantes sea un problema especialmente relevante, a pesar de ser uno de los explícitamente mencionados por Muñiz y Fonseca-Pedrero. Sin embargo, sí que podrían mencionarse otra serie de características. Por ejemplo, en lo relativo a la longitud, el consejo de no utilizar cuestionarios excesivamente largos para evitar el efecto Halo en los alumnos (Marsh, 1987, citado en Martínez 2005); o en lo relativo a las instrucciones, el papel que pueden jugar para reducir errores en la presentación de ítems no aplicables en algunos casos.

A continuación debe iniciarse el *Proceso de Construcción de Ítems*. De nuevo Muñiz y Fonseca-Pedrero se limitan a exponer algunas de las consideraciones generales que suelen encontrarse en la literatura relativa a la elaboración de ítems. Sin embargo, en esta ocasión incluyen un consejo que puede resultar problemático. Me refiero a su recomendación de que "los ítems deben de ser heterogéneos y variados para así recoger una mayor variabilidad y representatividad de la variable de medida" (p. 20). Si bien es cierta la relación entre variabilidad de los ítems y representación de la variable medida, la heterogeneidad de los ítems tiene su lado problemático. El autor del test debería, en algún momento, hacer explícito el modelo de medida desde el que trabaja. Una de las posibles alternativas es el modelo de Likert (1932). En este modelo se asume explícitamente que todos los ítems deben ser homogéneos y no aportar variabilidad por sí mismos a la medida. Quizá convenga recordar aquí que el modelo de Likert es algo más que un simple conjunto de ítems con respuesta graduada. En los cuestionarios de evaluación del profesorado utilizados en las universidades españolas, la respuesta graduada es claramente la norma. Ahora bien, la elevada variedad de ítems encontrada hace dudar seriamente de que esos mismos cuestionarios pudiesen cumplir los supuestos que el modelo de Likert conlleva.

El quinto apartado se etiqueta como *Producción, Base de Datos, Normas de Puntuación y Corrección*. Coincido plenamente con los profesores Muñiz y Fonseca-Pedrero en la importancia del apartado y en la consideración de que es habitualmente olvidado o minusvalorado. Pero coincidiendo con ellos en los argumentos que justifican la importancia del apartado, creo que se olvidan de mencionar una de las razones principales para cuidar la edición de esta primera versión del instrumento. El estudio piloto que se ha de realizar a continuación debe reproducir de la forma más fiel posible las condiciones reales en las que se llevará a cabo la evaluación. Una de estas condiciones es, precisamente, la naturaleza y forma del cuestionario. Pequeños cambios en la edición del cuestionario pueden provocar cambios mucho mayores en las respuestas de los participantes. Esta variabilidad no podrá ser controlada si ambas formas difieren considerablemente.

La sección dedicada a los *Estudios Cualitativo y Cuantitativo* presenta, como el resto, una serie de consideraciones generales que serían de utilidad en la elaboración de cualquier instrumento destinado a medir cualquier variable. No obstante, como hemos señalado reiteradamente, la evaluación de la docencia puede enfocarse desde perspectivas muy diferentes que llevarían a estrategias de análisis radicalmente distintas. A modo de ejemplo, dado que abordamos esta cuestión con más profundidad en otro apartado, podemos señalar que la variabilidad en las respuestas de los alumnos puede ser un buen indicador cuando ellos son el objeto de medida, pero se convierte en un indicador de error de medida cuando los alumnos son considerados el "instrumento" de medida de la actuación docente del profesor.

Considero que el nombre otorgado a la siguiente sección: *Selección de Otros Instrumentos de Medida Convergentes*, es equivocado e incompleto. Como bien señalan Muñiz y Fonseca-Pedrero la definición sintáctica es la fuente de la que se van a extraer las hipótesis sobre cómo deben relacionarse las puntuaciones obtenidas en el cuestionario con otras variables. Pero estas otras variables no incluyen sólo otros constructos con los que nuestra medida deba converger. Es posible, y deseable, la especificación de otras medidas divergentes, así como la relación con otro tipo de variables como criterios o variables sociodemográficas como la antigüedad en el puesto, categoría profesional, edad, etc. Todas estas relaciones formarán parte del análisis de la validez que Muñiz y Fonseca-Pedrero incluyen en el apartado posterior denominado estimación de las propiedades métricas. Considerando, por tanto, que los análisis relativos a la validez se realizarán posteriormente, este punto del proceso puede limitarse a la selección de todas las medidas que serán útiles en este cometido, y no sólo los instrumentos de medida convergentes.

La selección de la muestra es el primer asunto tratado en el punto rotulado como *Estudio de Campo*. Volviendo a lo señalado en reiteradas ocasiones, para especificar adecuadamente la muestra, deberemos aclarar previamente la población. ¿Son los profesores o los alumnos? ¿hay una población o varias? Además, y coincidiendo en las bondades generales del muestreo probabilístico creo conveniente un par de matizaciones. Una de las principales ventajas de las muestras probabilísticas es la de posibilitar la inferencia estadística. Sin embargo, en un estudio de campo dentro del proceso de elaboración de un cuestionario, la inferencia estadística no es el principal de los objetivos. Además, aunque el azar se pretende que actúe como garante de la participación en la muestra de toda una

serie de características no siempre es un aval suficiente. En la evaluación docente se pueden identificar toda una serie de variables con un efecto claro. No referimos a temas como la licenciatura, área de conocimiento, curso, etc. Asegurar explícitamente su representación en la muestra me parece fundamental, ya sea mediante un muestreo aleatorio estratificado o mediante un simple muestro por cuotas.

También dentro del estudio de campo plantean Muñiz y Fonseca-Pedrero que las condiciones físicas de la administración sean las adecuadas (en términos de luz, comodidad, ruidos, etc.). Esto sólo es cierto si las condiciones finales de administración reunirán las mismas facilidades. Si no es así, es mejor reproducir en el estudio de campo las mismas limitaciones con las que contará la administración definitiva.

El penúltimo apartado está dedicado a la *Estimación de las Propiedades Métricas*. Mencionan aquí los autores aspectos relativos al análisis de ítems, de la dimensionalidad, de la fiabilidad, de la validez y de la baremación. En relación a la dimensionalidad señalan que "en el caso de encontrar una solución esencialmente unidimensional nos podríamos plantear la construcción de una puntuación total, en el caso de una estructura multidimensional deberíamos pensar en un conjunto de escalas o perfil de puntuaciones" (p. 22). Creo que esta afirmación puede llevar a error. La estructura dimensional del cuestionario no es algo que debamos encontrar en este momento del proceso. Por el contrario, debe haber sido especificada con antelación. Este es el momento, dado que es cuando por primera vez contamos con datos para ello, de confirmar que la dimensionalidad de la prueba coincide con lo previsto. De igual modo, la forma de puntuación ya debía estar especificada en un punto anterior.

En el análisis de la fiabilidad Muñiz y Fonseca-Pedrero plantean las alternativas de la teoría clásica de los test y de la teoría de respuesta a los ítems. En este sentido me gustaría hacer dos comentarios. En primer lugar un recordatorio de que también es posible abordar el estudio de la precisión desde la perspectiva de la teoría de la generalizabilidad (Cronbach, Gleser, Nanda y Rajaratnam, 1972; Brennan 2001). Este procedimiento que, en cierta medida puede ser considerado una extensión de la TCT, presenta la ventaja de posibilitar la importancia de más fuentes de error distintas al aleatorio. Por otro lado, estas opciones son adecuadas únicamente si se está contemplando una perspectiva de escalamiento centrada en la persona. Si el marco conceptual encaja mejor en una perspectiva centrada en los estímulos, serían más aconsejables análisis de la precisión encaminados a estimar el acuerdo entre jueces. En cualquier caso, la estimación del coeficiente alfa como estimador de la fiabilidad plantearía serios inconvenientes. Si bien podría ser aceptable la estimación de distintos coeficientes alfa para las distintas dimensiones que cubren los cuestionarios; sería necesaria una reducción previa de los datos. No es adecuado calcularlo directamente sobre la matriz rectangular resultante de codificar las respuestas de todos los alumnos a los cuestionarios. Siguiendo la lógica habitual de codificación, en este caso cada fila correspondería a un alumno, mientras que en el cálculo del coeficiente alfa cada fila debe corresponder a una persona evaluada (en este caso profesores).

Con lo que respecta a la validez, Muñiz y Fonseca-Pedrero sólo citan la relación del instrumento de medida con otros instrumentos de evaluación. Ya hemos comentado anteriormente que la red de relaciones de la definición sintáctica puede ser más amplia que esto. Ahora nos gustaría señalar que las posibilidades de los estudios de validación incluyen otros aspectos que, especialmente en este contexto, pueden resultar relevantes. Nos referimos al análisis de los procesos de respuesta como fuente de evidencias para la validez (American Educational Research Association, American Psychological Association y National Council on Measurement in Education, 1999). Ya anticipamos esta posibilidad en la discusión del apartado dedicado a los estudios cualitativos. Este tipo de estudios pueden considerarse como evidencia de validez, y es por lo que los abordamos aquí. Los sesgos de respuesta o el efecto halo, entre otros, pueden ser fuente de varianza irrelevante para el constructo (Messick, 1989) que invaliden la interpretación prevista de las puntuaciones. Este tipo de evidencia no puede encontrarse a partir, exclusivamente, del análisis de las relaciones del cuestionario con otros instrumentos de evaluación.

Para terminar el apartado del estudio de campo, Muñiz y Fonseca-Pedrero instan a realizar una baremación del instrumento de medida donde se establecen puntos de corte normativos. Esta baremación tiene sentido cuando en un futuro serán posibles administraciones individuales del instrumento y se desea contar con un marco con el que comparar e interpretar esas puntuaciones. Pero ese no es el caso habitual de los estudios de evaluación docente. La realización final de esas

evaluaciones implicará, en muchos casos, la medida de prácticamente toda la población de interés. Será más correcto e informativo establecer los baremos necesarios sobre esa población definitiva que sobre la muestra menor que participa en el estudio de campo.

El último apartado del proceso es el dedicado a la *Elaboración de la Versión Definitiva, Informe y Manual del Cuestionario*. Nada que añadir en este punto, tan sólo recordar que será un reflejo de lo anteriormente desarrollado, con sus bondades y flaquezas.

Las conclusiones a las que llegan Muñiz y Fonseca-Pedrero destacan en primer lugar la complejidad de los procesos de evaluación universitaria –aspecto en el que estoy completamente de acuerdo–. Además señalan que tales procesos deben responder a toda una serie de cuestiones. Los autores justifican, de forma razonable, la poca profundidad en la que se aborda el proceso de construcción en términos de las limitaciones de espacio. Si bien es cierto que el tratamiento en profundidad de los aspectos relativos a la elaboración de un cuestionario para la evaluación docente excedería de las posibilidades de un artículo clásico, creo que la opción de simplificar lo sobradamente conocido es una alternativa errónea. Las particularidades de la evaluación universitaria pueden hacer que decisiones que serían generalmente correctas en el terreno de la elaboración de test se conviertan en acciones equivocadas.

4. CONCLUSIONES

No pretendo negar la alta importancia que puede tener la evaluación de la actuación docente del profesorado por parte de los alumnos. Pero el esquema de una serie de alumnos respondiendo a un conjunto de preguntas no puede ser la panacea que sirva para cualquier propósito. Podemos encontrar distintos escenarios en los que la utilidad de esa forma de proceder varía sustancialmente. Por un lado, la situación de medida descrita puede ser el reflejo de un interés en conocer algo sobre los alumnos. Por ejemplo, su satisfacción o su valoración con determinados aspectos de la actuación docente de un profesor. En este caso, las características relevantes del proceso serían, entre otras:

1. La población objetivo de la medida son los alumnos.
2. Las preguntas del cuestionario pueden, o no, combinarse para obtener otros indicadores.
3. La variabilidad en las repuestas de un mismo alumno a distintas preguntas del cuestionario no es necesariamente buena ni mala.
4. La variabilidad en las respuestas de distintos alumnos a una misma pregunta no puede considerarse negativa, sino simple reflejo de distintas opiniones o valoraciones por su parte.
5. Lo más importante, los ítems del cuestionario han de hacer referencia, exclusivamente, a aspectos sobre los que el alumno deba y pueda indicar su satisfacción.

En definitiva, en este caso, la fotografía del conjunto de alumnos respondiendo a un cuestionario refleja claramente una situación de encuesta. Por tanto, cobran especial interés todos los problemas metodológicos asociados a esta situación. Los errores de muestreo y no muestreo, la inferencia estadística, el tratamiento de los datos, etc. No olvidemos que en esta situación, raramente interesará la opinión de un alumno en particular sobre un profesor en particular.

Un aspecto muy importante de este escenario es que, cuando se realiza en el conjunto de una Universidad no se puede concebir el conjunto de los cuestionarios como formando parte de una misma encuesta.

En una encuesta hay una población claramente definida y un objeto de medida o variable, cuyo valor se quiere conocer en esa población. Pero ¿cuál es la variable que hay que estimar en las encuestas a alumnos? No existe nada parecido a la "actuación docente del profesor tipo". Cada vez que se cambia la clase o grupo de alumnos cambia la variable medida y se hace imposible unificar los datos. Podemos comparar la situación con una investigación a nivel nacional donde se pide a los entrevistados que valoren algún atributo del alcalde de su municipio. En realidad se están realizando tantas encuestas como municipios, y no tiene sentido combinar las respuestas de habitantes de ciudades distintas que valoran alcaldes distintos. Es más, dado que se trata de poblaciones distintas, con variables de interés distintas, debería plantearse la conveniencia de adaptar los cuestionarios para optimizar su funcionamiento en cada población.

La situación es completamente distinta de una en la que se preguntase a todas esas personas por una misma variable. Por ejemplo, su valoración del presidente de gobierno. En este caso, cada municipio

es simplemente una subpoblación sobre la que es posible extraer inferencias particulares y el total de la muestra puede utilizarse sin problemas para realizar inferencias al total de la población.

Pero la fotografía antes descrita de un grupo de alumnos respondiendo a un cuestionario con preguntas sobre su profesor puede ser el retrato de una situación completamente distinta. Una en la que se invierten los papeles y los alumnos se convierten en el instrumento de medida, mientras que los profesores pasan a ser el objeto de la misma.

Cabe imaginar la utilidad de conocer la "actuación docente de los profesores" mediante cuestionarios que se administran a los alumnos. Pero en esta situación habría que hacer una serie de precisiones. Esto tendría sentido si se llega a la conclusión de que preguntar a los alumnos es la mejor manera de obtener un indicador de la conducta que lleva a cabo el profesor y que es un reflejo de su nivel en la variable psicológica. Como ya he señalado, en esta situación los alumnos jugarían el papel de instrumentos de medida, de jueces, y no de objetos de medida. Este diferente papel tiene una consecuencia inmediata. Ya no es necesario, ni siquiera tiene por qué ser conveniente, administrar el cuestionario a todos los alumnos o a una muestra representativa de ellos. Además, su opinión ya no importa, importa sólo su juicio. Si la diversidad de opiniones es siempre respetable, la diversidad de juicios sobre un mismo objeto no puede ser catalogada de otra forma que de error. Siguiendo con esta suposición de que los alumnos son el instrumento de medida, el siguiente paso sería determinar la forma de corrección del ítem. Hay distintas posibilidades, puede optarse por la mediana o bien por la moda. Por la menor o mayor categoría que alcance una determinada frecuencia relativa, etc. O como señala Morales (2006) específicamente en los cuestionarios de evaluación del profesorado "si queremos hacer un estudio sobre cómo los alumnos evalúan a sus profesores, la unidad de análisis, el sujeto estudiado, es en principio el profesor, no el alumno" (Morales, 2006, p.3) lo que lleva a este autor a aconsejar claramente el empleo de medias grupales en lugar de puntuaciones individuales de los alumnos.

Pero el hecho de que preguntar a los alumnos sea una forma aceptable a priori de evaluar a los profesores no la convierte en la más adecuada. La obtención de muestras de conducta mediante informantes indirectos, o "proxies" aunque poco habitual, no es nueva en la evaluación psicométrica. Una posibilidad son las situaciones donde las personas objeto de medida no están capacitadas para informar directamente sobre su comportamiento. Por ejemplo, el cuestionario LittleEars de desarrollo auditivo (Kuehn-Inacker, Weichbold, Tsiakpini, Coninx y D'Haese, 2003). Pero en estos casos es necesario justificar adecuadamente el porqué salirse de la norma y optar por una fuente indirecta. Si la evaluación universitaria se realiza de esta forma, sería conveniente señalar las razones que llevan a elegir a los alumnos como fuente de información sobre la conducta de los profesores en lugar de interrogar directamente a éstos.

En definitiva, parece que existe un consenso generalizado tanto en el reconocimiento de la importancia de la evaluación universitaria como en la forma de abordarla. En este trabajo he intentado mostrar mi acuerdo con el primero de estos puntos y mi desacuerdo con el segundo. Para ello he utilizado como excusa el trabajo de Muñiz y Fonseca-Pedrero, si bien considero que mis comentarios son aplicables a la mayoría de procesos de evaluación universitaria que se realizan en las universidades españolas. Básicamente, la crítica fundamental se refiere a la confusión e indefinición del papel que juegan los distintos agentes implicados.

Muñiz y Fonseca-Pedrero finalizan su trabajo con un llamamiento a que estos procesos de evaluación se realicen por profesionales cualificados y no se cometa la temeridad de dejarlos en manos de aficionados bienintencionados. Me gustaría poder compartir esta opinión.

BIBLIOGRAFÍA

American Educational Research Association, American Psychological Association, y

National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

Brennan, R. L. (2001). Generalizability theory. New York: Springer-Verlag.

Crocker , L. y Algina, J. (1986). *Introduction to Classical & Modern Test Theory*. New York : Holt, Rinehart and Winston

Cronbach, L. J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley & Sons, Inc.

Defior , S.; González, A.; Gutierrez, R.;. Gutierrez, J.; Maldonado, J.A.; Montabes, J.; Núñez, J.; Padilla, J.L.; Pascual, A.; Pino, J.L.; Rico, L.; Rufian, A.; Torres, F. y Vela, M. (2007). *Opinión del Alumnado Sobre la Actuación Docente del Profesorado de la Universidad de Granada: Resultados 2004-2007*. Granada, España. Editorial Universidad de Granada. 2007. 730.

Likert,R. (1932) A technique for measurement attitudes. *Archives of Psychology*. n. 140.

Kuehn-Inacker, H.; Weichbold, V.; Tsiakpini, L.; Coninx, F.y D'Haese, P. (2003) *LittlEARS Auditory Questionnaire Manual – Parent questionnaire to assess auditory behaviour in young children.* Innsbruck, Austria: MED-EL

Marsh, H.W. (1987). Students evaluations of university teaching: Research, findings, methodological issues and directions for future research. *International Journal of Education Research*, 11, (3), 253-388.

Martínez, M (2005) *Estudio del cuestionario de evaluación del Profesorado de la upv mediante opinión de los Estudiantes*. Tratamiento estadístico. Tesis doctoral: Universidad de Valencia.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). New York: American Council on Education.

Morales, P. (2006) *El problema de la unidad de análisis en la investigación educacional: datos individuales o medias de grupos*. Documento web. URL: http://www.upcomillas.es/personal/peter/investigacion/UnidadAnalisis.pdf. FEcha de acceso: mayo 2009.

Muñiz, J. y Fonseca-Pedrero E. (2008) Construcción de instrumentos de medida para la evaluación universitaria. *Revista de Investigación en Educación*, nº 5, 2008, pp. 13-25.

Muñoz Cantero, J.M., Ríos de Deus, M.P y Abalde, E. (2002). Evaluación Docente vs. Evaluación de la Calidad. *Revista ELectrónica de Investigación y Evaluación Educativa* (RELIEVE), v. 8, n. 2, 103-134. http://www.uv.es/RELIEVE/v8n2/RELIEVEv8n2_4.htm. Consultado en (mayo 2009).

Torgenson, W.S. (1958). *Theory and Methods of Scaling*. New York: John Wiley and Sons, Inc. Cuarta Impresión, 1963.