ARTÍCULO ORIGINAL

CONSTRUCCIÓN DE INSTRUMENTOS DE MEDIDA PARA LA EVALUACIÓN UNIVERSITARIA

JOSÉ MUÑIZ y EDUARDO FONSECA-PEDRERO
CIBERSAM, Universidad de Oviedo

RESUMEN: En el artículo se analizan los problemas metodológicos implicados en la evaluación universitaria. Se trata de analizar todos aquellos aspectos técnicos y de procedimiento que son necesarios para llevar a cabo evaluaciones rigurosas en el ámbito universitario, tanto en lo relativo a investigación, como a la docencia y gestión. Para ello se repasarán los distintos componentes de un proceso evaluativo integral, a saber, qué se evalúa, partes legítimamente implicadas, quién evalúa, cómo se evalúa, feedback a las partes, planes de mejora, y opinión de las partes. Se hará especial hincapié en los requisitos metodológicos necesarios para que los instrumentos de medida utilizados en la evaluación sean fiables y válidos. Para ello se describen diez pasos necesarios que hay que seguir para construir y analizar este tipo de instrumentos de medida. Se finaliza discutiendo las perspectivas de futuro en el ámbito de la evaluación universitaria.

PALABRAS CLAVE: Evaluación, Universidades. Construcción de tests. Escalas.

ABSTRACT: In this paper, we analyze the methodological problems involved in the evaluation of universities. We describe the technical and procedural aspects required to carry out rigorous assessments in the college context, particularly with respect to research and teaching. We start by reviewing the different components of the integral evaluation process, such as determining what is assessed, who are the assessors, how to asses in a technically sound way, providing feedback to all stakeholders, developing improvement plans derived from the evaluation, and measuring the opinions of those involved in the evaluation process. Special attention is paid to the technology and the methodology required for developing reliable and valid assessment instruments. Ten steps followed to develop rigorous and objective assessment instruments are described in detail, pointing out the possible difficulties and problems to be found. Finally, we discuss future directions for college evaluation.

KEY WORDS: Evaluation. University. Test development. Scaling.

1. INTRODUCCIÓN

Parafraseando al clásico bien podría decirse que un fantasma recorre la Universidad Española, es el fantasma de la evaluación, de repente todo el mundo se ha puesto a evaluar a todo el mundo a todas horas.

Fecha de recepción 08-10-2008 · Fecha de aceptación 17-12-2008
Correspondencia : José Muníz
CIBERSAN, Universidad de Oviedo

Tal vorágine evaluativa tiene como finalidad legítima la elaboración de un diagnóstico riguroso que permita mejorar los tres grandes parámetros que determinan la calidad de una Universidad: Investigación, Docencia y Gestión. Ciertamente, sin una evaluación precisa de esos tres parámetros no se puede hacer un diagnóstico riguroso y certero que permita generar planes de mejora basados en datos empíricos. Los rankings de universidades, tanto nacionales como internacionales ejercen una fuerte presión y nadie quiere quedarse atrás (Buela-Casal, Bermúdez, Sierra, Quevedo-Blasco y Castro-Vázquez, 2009).

Desde un punto de vista metodológico, una evaluación integral de cualquier organización o institución como es la Universidad requiere

disponer de un modelo general de evaluación que integre y dé sentido a las distintas evaluaciones específicas que necesariamente se llevarán a cabo. Más allá de algunas deficiencias y limitaciones técnicas de las evaluaciones concretas realizadas, seguramente la limitación estructural más importante en la evaluación actual de las universidades es la carencia de un modelo general que integre y guíe las evaluaciones sectoriales. Este modelo general debe de dar respuesta clara y operativa al menos a siete cuestiones clave:

- 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
- opinión de las partes implicadas sobre la evaluación.

A continuación se comentan de forma somera los problemas implicados en cada una de estas facetas, para finalmente centrarnos en lo que constituye el núcleo central de este artículo, a saber, cómo se evalúa, qué propiedades métricas deben de tener los instrumentos de evaluación universitaria para obtener datos fiables y válidos que sirvan para tomar decisiones fundadas. Conviene dejar claro ya desde el principio que disponer de instrumentos de evaluación técnicamente adecuados no garantiza en absoluto un proceso evaluativo global exitoso, es una condición necesaria, pero no suficiente. Si se dispone de excelentes instrumentos de medida desde el punto de vista métrico, pero se descuida alguno de los siete aspectos citados del modelo integral de evaluación habremos fracasado en el proceso de evaluación.

1.1. Qué se evalúa

Lo primero y fundamental que hay que hacer cuando se planifica una evaluación es definir de forma clara, concisa y operativa aquello que se desea evaluar. Esto parece obvio y de sentido común, pero a menudo los objetivos de la evaluación aparecen confusos y pobremente definidos, con lo cual la evaluación está condenada al fracaso, utilícese la metodología que se utilice. Una definición operativa, es decir, susceptible de ser medida, obliga a buscar un compromiso entre la riqueza del constructo a medir y la objetividad de los instrumentos de medida utilizados. Tan vano es plantearse grandes objetivos que son imposibles de medir de forma rigurosa, como medir de forma muy precisa aquello que es irrelevante pero muy medible. La virtud está en el punto medio, hay que llegar a un compromiso para evaluar lo esencial del constructo, y hacerlo de forma válida y fiable, lo uno sin lo otro conduce a una evaluación fracasada.

En el contexto de la evaluación universitaria existen numerosos objetivos de evaluación, pero casi todos ellos se pueden clasificar en tres grandes bloques: individuos, productos y sistemas. En el caso de los individuos se evalúan alumnos, profesores, personal de administración y servicios, o gestores (Rectores, Decanos, Directores de Departamento, etc.). La evaluación de cada uno de estos aspectos conlleva una problemática específica en la que resulta imposible entrar aquí, véase, por ejemplo Centra (1993), Fernández (2008), Fernández, Mateo y Muñiz (1995, 1996), Aleamoni (1999), Beran y Violato (2005) y Marsh y Roche (2000) para todo lo relativo a la evaluación del profesorado. Conviene llamar la atención sobre la escasa formación que los profesores universitarios suelen tener sobre la metodología de la evaluación educativa, imprescindible para llevar a cabo una evaluación objetiva de los estudiantes (Brennan, 2006; *Joint Committee on Standards for Educational Evaluation*, 2003). En cuanto a los productos evaluados la gama es ciertamente amplia, destacando los proyectos de investigación, artículos, tesis, libros, curricula, planes de estudios, etc. En cuanto a los sistemas evaluados pueden citarse los departamentos, facultades, institutos, grupos de investigación, másters, bibliotecas, y universidades como tales.

Como se puede ver las dimensiones para evaluar en el contexto universitario son muchas y variadas, si bien no conviene que los árboles nos impidan ver el bosque, los dos grandes parámetros que determinan la calidad de una Universidad son, por un lado, la calidad de los alumnos, y por otro, la calidad de los profesores. La calidad de una Universidad viene dada por el producto de esos dos factores, igual que el rendimiento de un alumno viene dado por el producto de su capacidad por su esfuerzo. Atraer a los mejores alumnos y a los mejores profesores es lo que garantiza la calidad de una Universidad, si bien hay otros factores complementarios de interés, tales como el número de alumnos por profesor, las bibliotecas, las facilidades informáticas, el tamaño de la Universidad, la gestión realizada, la calidad de aulas, laboratorios, instalaciones deportivas, cultura de empresa, y un largo etcétera. La mayoría de los

rankings que se hacen sobre las universidades (*International Ranking Expert Group*, 2006) miden de un modo u otro esos dos grandes factores para establecer las clasificaciones. Así, por ejemplo, el popular ranking elaborado por el diario británico *The Times* asigna un 60% a la calidad de la investigación, un 10% a la capacidad de que un graduado encuentre trabajo, otro 10% a la presencia internacional, y un 20% a la relación número de estudiantes-profesores. Por su parte el bien conocido ranking de Shanghai (*Institute of Higher Education*, 2008) pondera con un 10% el número de premios Nobel, con un 20% a los ganadores de la Medalla Fields (una especie de Nobel de Matemáticas), 20% a los investigadores altamente citados, 20% a los artículos publicados en las revistas *Nature* y *Science*, 20% al impacto de los trabajos registrados por *Science Citation Index*, y finalmente un 10% al tamaño de la institución, que sí cuenta. Ni que decir tiene que las universidades españolas no aparecen entre los cien primeros puestos de estas clasificaciones, ni están en vías de aparecer. El análisis del porqué de esta situación nos llevaría lejos de los objetivos de este trabajo, pero las causas profundas hay que buscarlas en un sistema que impide que las universidades seleccionen a los mejores alumnos y a los mejores profesores e investigadores, así de simple y así de complejo. En un trabajo reciente Buela-Casal et al. (2009) encuentran que los profesores de Universidad españoles asignan los siguientes valores (de 1 a 5) a los criterios para evaluar la producción científica: Artículos en revistas con Factor de Impacto (4,19), Tramos de Investigación (3,95), Proyectos I+D conseguidos (3,90), Tesis doctorales dirigidas (3,47), Becas FPU (3,03), y Doctorados con mención de calidad (3,02). Estos valores dan una idea bastante clara de lo que piensa la comunidad universitaria acerca de la producción científica de los investigadores.

1.2. Partes legítimamente implicadas en la evaluación

Si bien en cada caso concreto pueden existir ligeras variantes, en el contexto universitario los agentes legítimamente implicados en las evaluaciones son entre otros: los alumnos, los padres de éstos, los profesores, el personal de administración y servicios, los gestores universitarios, y la propia sociedad, representada por los políticos elegidos, que subvenciona la Universidad. Se olvida con demasiada frecuencia que lo aportado por las matrículas del alumno apenas si cubre un diez por ciento de lo que cuesta su formación, siendo aportado el resto por los impuestos de los ciudadanos, tengan o no hijos en la Universidad. A la hora de llevar a cabo cualquier evaluación debe definirse con precisión qué papel juega cada uno de estos agentes implicados, el cual variará en función de la naturaleza y fines de la evaluación.

1.3. Quién evalúa

Según la relación de los agentes evaluadores con la institución, suele hablarse de evaluación externa, cuando los evaluadores, personas o agencias, son externos a la institución, evaluación interna, cuando pertenecen a la propia institución evaluada, o mixta, si es una mezcla de ambas. No existe una regla universal, los tres modelos son legítimos y dependerá de cada caso que se elija un modelo u otro. Así, es habitual que las propias universidades evalúen la actividad docente de los profesores, si bien en algunos casos se recurre a instancias externas. Sin embargo, para evaluar los proyectos de investigación y los curricula suele recurrirse a agencias externas, o utilizar modelos mixtos. Una evaluación no es mejor ni peor por ser realizada externa o internamente, de lo que se trata es que sea objetiva, rigurosa, independiente, fiable y válida. Cada caso aconsejará si para obtener estos resultados es mejor recurrir a evaluaciones externas, internas o mixtas.

1.4. Cómo se evalúa: qué metodología utilizar

Desde un punto de vista métrico, los instrumentos que se utilicen para la evaluación han de ser objetivos, claros, comprensibles por las partes, preferiblemente cuantitativos, fiables y válidos. Más adelante detallaremos los pasos que deben seguirse para desarrollar instrumentos de evaluación que cumplan estas condiciones. No obstante, conviene aclarar desde el principio que un instrumento métricamente adecuado es condición necesaria, pero no suficiente para llevar a cabo una evaluación exitosa (Muñiz y Bartram, 2007). Aspectos como el proceso de aplicación del instrumento, el uso que se hace de los resultados, o el *feedback* dados a las partes, pueden hacer que un excelente instrumento no genere los resultados deseados de la evaluación. En suma, la evaluación es un proceso, uno de cuyos componentes son los instrumentos de evaluación, pero no los únicos.

1.5. *Feedback* a las partes implicadas

La finalidad de toda evaluación universitaria es generar mejoras en la calidad universitaria, y para que ello se produzca es fundamental dar el *feedback* adecuado a las partes implicadas. Una evaluación técnicamente perfecta no cumpliría su objetivo si no se hiciera llegar la información correspondiente a las partes legítimamente implicadas. Aparte de los planes de mejora que se elaboren a partir de la evaluación, el mero hecho de comunicar de forma adecuada los resultados constituye el agente de cambio más eficaz y económico. Meter los resultados de una evaluación en un cajón, o no hacer llegar a quien corresponda el *feedback* pertinente es un error que debe evitarse. La explicación y difusión de los resultados constituye una parte esencial del proceso evaluador. Debe de analizarse y estudiarse con suma precisión y rigor cuál es la mejor manera de dar el *feedback* a las partes implicadas, buscando la máxima efectividad, en el sentido de generar mejoras en el sistema. Por ejemplo, ¿cómo se deben de dar a los profesores los resultados de la evaluación hecha por los estudiantes? ¿Han de hacerse públicos en el tablón de anuncios del centro? ¿Han de ser privados? ¿Deben conocerlos el decano y el director del departamento? ¿Qué información debe de incluirse? De poco vale que el cuestionario utilizado sea excelente si luego se falla a la hora de solucionar estas y otras cuestiones (Fernández, 2008). No hay reglas específicas que resuelvan toda la casuística evaluativa universitaria, pero sí dos pautas generales que hay que seguir a la hora de generar el *feedback*. En primer lugar, el *feedback* ha de ser rápido, cuanto más tiempo pase entre la evaluación y el momento de proporcionar *feedback* más ineficaz se vuelve la evaluación. Esto vale para todo tipo de evaluaciones, y en especial para las calificaciones de los alumnos. En segundo lugar, el *feedback* ha de ser conciso y claro de interpretar, ajustando su formato e información en función del destinatario. Antes de llevar a cabo la evaluación ha de preverse exactamente el *feedback* que se va a proporcionar.

1.6. Planes de mejora

Toda evaluación está destinada a terminar en un plan de mejora de algún aspecto institucional. Existen planes de mejora de todo tipo y condición, si bien todos ellos deben reunir determinados requisitos. En primer lugar, han de estar fundamentados en datos objetivos de partida y tener también metas objetivas de llegada, que sean evaluables preferiblemente de forma cuantitativa. Han de ser acordados por las partes implicadas, si un plan de mejora no está consensuado con las partes es altamente probable que no funcione. Han de formularse de forma objetiva, contemplándose en el diseño la forma de evaluarlos, es decir, la evaluación forma parte del propio plan de mejora. Un plan no evaluable objetivamente en realidad no es un plan, es como mucho un deseo bienintencionado. Hay que evitar generalizaciones hueras del tipo: el plan propuesto se propone re-culturalizar la Facultad. Si un plan de mejora no conlleva el diseño de su evaluación objetiva no puede considerarse como tal en sentido estricto. Es recomendable llevar a cabo una evaluación previa al plan, una posterior al plan, y hacer un seguimiento a medio y largo plazo. Siempre que sea posible es recomendable utilizar uno o más grupos de control.

1.7. Opinión de las partes implicadas en la evaluación

El proceso de evaluación termina con la recogida de información de las partes implicadas a los distintos niveles de la evaluación. Este aspecto es fundamental, pues va a permitir llevar a cabo reformulaciones y ajustes cara a futuros planes de evaluación. La información puede recogerse de muy diversas formas, incluyendo encuestas de opinión, cuestionarios, reuniones de grupo, entrevistas personales, etc. Se trata de hacer una recogida de información lo más objetiva, rigurosa, fiable y válida posible. Aparte del valor intrínseco de la información recogida, se fomentará la identificación e implicación de los distintos agentes en el proceso, sintiéndolo más suyo y cercano.

Una vez comentados los siete aspectos fundamentales implicados en el proceso de evaluación universitaria, vamos a centrarnos ahora en los distintos pasos y actividades que habría que seguir para desarrollar instrumentos de medida con unas propiedades métricas deseables para su empleo en la evaluación universitaria.

2. CONSTRUCCIÓN DE INSTRUMENTOS DE MEDIDA PARA LA EVALUACIÓN UNIVERSITARIA

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. La importancia de unos instrumentos de evaluación adecuados radica en la trascendencia de las decisiones y las consecuencias que a partir de ellos se derivan, tanto personales como sociales

(Anastasi y Urbina, 1998; Kane, 2006; Messick, 1998; Muñiz, 1997b; Padilla, Gómez, Hidalgo y Muñiz, 2006; Padilla, Gómez, Hidalgo y Muñiz, 2007; Sireci, 2007; Sireci y Parker, 2006; Zumbo, 2007). Si el proceso de construcción se lleva cabo de forma defectuosa las inferencias que se obtengan a partir de las puntuaciones y la toma de decisiones que de ellas se deriven serán totalmente equivocadas e infundadas (Elosúa, 2003; Muñiz, 2004; Muñiz, Fidalgo, García-Cueto, Martínez y Moreno, 2005; Schmeiser y Welch, 2006).

Los requisitos técnicos que debe cumplir un instrumento de evaluación aparecen bien documentados en la literatura especializada (*American Educational Research Association, American Psychological Association y National Council on Measurement in Education*, 1999; Carretero-Dios y Pérez, 2005; Clark y Watson, 1995; Downing, 2006; Morales, Urosa y Blanco, 2003; Muñiz, 1996, 1997a, 2000; Nunnally y Bernstein, 1995; Schmeiser y Welch, 2006; Smith, Fischer y Fister, 2003; Wilson, 2005). La construcción de un instrumento de medida es un proceso complejo que se puede articular en varios pasos, si bien éstos no son automáticos y universales, pudiendo variar en función del propósito del instrumento de medida (selección, diagnóstico, etc.), del tipo de respuesta (selección o construcción), del formato de administración (lápiz y papel o informatizado), o del contexto de evaluación (exámenes, evaluación docente, etc.), por citar sólo algunos casos. Todo el proceso de construcción debe ser definido objetivamente siguiendo unos principios teóricos y métricos para así maximizar su validez (Downing, 2006; Smith, 2005). Puede decirse que el proceso de validación ya comienza a fraguarse antes de la propia elaboración del instrumento, pues todas las acciones que realicemos antes, durante y después permitirán recoger datos empíricos que ayuden a la interpretación de las puntuaciones (Elosúa, 2003; Muñiz, 2004; Zumbo, 2007).

3. PASOS PARA LA CONSTRUCCIÓN DE UN INSTRUMENTO DE MEDIDA

En la Tabla 1 se recogen de forma esquemática las principales fases que se deben considerar en el proceso de construcción y validación de los instrumentos de medida, y a continuación se comenta cada una de ellas.

3.1. Marco general del instrumento de medida

Todo proceso de construcción de un instrumento de medida comienza por una justificación detallada y precisa de cuáles son las causas que motivan su construcción. Asimismo, hay que delimitar con claridad cuál es la variable objeto de medición, cuál va a ser el contexto de aplicación o circunstancias en el que se va a administrar el instrumento de evaluación, el tipo de aplicación (individual, colectiva), el formato de aplicación (lápiz y papel, informática), y qué decisiones se van a tomar a partir de las puntuaciones. Las causas que pueden llevar a la construcción de un instrumento de evaluación son diversas. Por ejemplo, un profesor universitario puede decidir construir un nuevo instrumento porque no existe ningún otro para medir una determinada variable, porque imparte docencia en una materia nueva y necesita evaluar a sus estudiantes, o simplemente porque los instrumentos existentes en el mercado presentan unas pésimas propiedades métricas. Los responsables de la construcción del instrumento de medida no sólo deben especificar el motivo por el cual quieren desarrollar un instrumento nuevo, sino también deben delimitar con claridad cuál es el contexto en el que se va a aplicar, lo que incluye necesariamente la población objeto de medición (alumnos, profesores, departamentos, etc.) y las circunstancias de aplicación (lugar, medios de los que se dispone y condiciones de aplicación, individual o colectiva). También debe de especificarse de antemano con qué propósito van a ser utilizadas las puntuaciones y qué decisiones se van a tomar a partir de ellas. En este sentido, las puntuaciones en un instrumento de evaluación pueden servir para propósitos varios como por ejemplo: seleccionar, diagnosticar, clasificar, orientar, evaluar un dominio específico o incluso como método de *screening* (*American Educational Research Association* et al., 1999). Se debe dejar claro que las inferencias que se extraigan de las puntuaciones de un instrumento de medida son siempre para un uso, contexto y población determinada. Así, lo que pueda ser válido para un grupo determinado de personas o población tal vez no lo sea para otra, y lo que pueda ser válido en un contexto de evaluación no tiene por qué serlo en otro diferente (Zumbo, 2007).

1. Marco general del instrumento de medida
- Justificación y motivación
- Contexto de aplicación
- Uso e interpretación de las puntuaciones

2. Definición operativa de la variable medida
- Definición operativa
- Definición sintáctica y semántica

3. Especificaciones del instrumento de medida
- Requerimientos de administración
- Tipo, número, longitud, formato, contenido y distribución de los ítems
- Especificaciones e instrucciones en la entrega del material
- Aspectos de seguridad

4. Construcción de los ítems
- Directrices para la construcción de ítems de elección múltiple
- Principios generales para la construcción de ítems

5. Producción, base de datos, normas de puntuación y corrección
- Composición
- Edición
- Puntuación y corrección

6. Estudio piloto cualitativo y cuantitativo
- Selección de la muestra piloto (cualitativo y cuantitativo)
- Análisis y resultados del estudio piloto (cualitativo y cuantitativo)
- Depuración, revisión, modificación o construcción de ítems
- Producción de una nueva versión del instrumento de medida

7. Selección de otros instrumentos de medida convergentes
- Obtener información convergente
- Utilizar pruebas ya validadas

8. **Estudio de campo**
- Selección y tamaño de la muestra y tipo de muestreo
- Administración del instrumento de medida
- Control de calidad y seguridad de la base de datos

9. **Estimación de las propiedades métricas**
- Análisis de ítems (cualitativo y cuantitativo)
- Dimensionalidad
- Estimación de la fiabilidad
- Obtención de evidencias de validez
- Tipificación

10.Versión definitiva, informe final y manual del instrumento de medida
- Prueba fina propuesta
- Manual

Tabla 1. Fases generales del proceso de construcción de instrumentos de medida

3.2. Definición operativa de la variable medida

El objetivo esencial de esta segunda fase es la definición operativa, semántica y sintáctica de la variable medida, así como las facetas o dimensiones que la componen (*American Educational Research Association* et al., 1999; Carretero-Dios y Pérez, 2005; Lord y Novick, 1968; Wilson, 2005).

El constructo evaluado debe definirse en términos operativos, para que pueda ser medido de forma empírica (Muñiz, 2004). En este sentido, tan interesante puede ser definir cuidadosamente lo que es el constructo como lo que no es. La facilidad o dificultad de la definición operativa depende en cierta medida de la naturaleza de variable objeto de medición.

Para llevar a cabo una definición operativa de la variable que nos interesa medir es clave realizar una revisión exhaustiva de la literatura publicada al respecto, así como la consulta a expertos (Clark y Watson, 1995; Wilson, 2005). Ello permite, por un lado, delimitar la variable objeto de medición, y considerar todas las dimensiones relevantes de la misma, y por otro, identificar con claridad los comportamientos más representativos de la variable de medición (Dolores y Padilla, 2004; Smith, 2005). Hay que evitar el dejar fuera alguna característica o dominio relevante del constructo (infraestimación), así como ponderar en demasía una faceta o dominio (sobreestimación) (Smith et al., 2003). Una definición operativa y precisa del constructo influye de forma determinante en la posterior obtención de los diferentes tipos de evidencias, ayuda a especificar las conductas más representativas de la variable objeto de medición y facilita el proceso de construcción de ítems (Carretero-Dios y Pérez, 2005; Elosúa, 2003; Muñiz et al., 2005; Sireci, 1998; Smith, 2005).

No sólo es importante una definición operativa de la variable sino que también es preciso identificar y definir las facetas o dominios del mismo (definición semántica) y la relación que se establece entre ellas así como con otras variables de interés (definición sintáctica) (Lord y Novick, 1968). La variable objeto de medición no se encuentra aislada en el mundo, sino que está en relación o interacción (positiva y/o negativa) con otras variables. Es interesante comprender y analizar estas relaciones especificándolas de antemano con el propósito de llevar a cabo posteriores estudios dirigidos a la obtención de evidencias de validez (Carretero-Dios y Pérez, 2005; Muñiz, 2004) y validación de teorías (Smith, 2005).

3.3. Especificaciones del instrumento de medida

Una vez delimitados el propósito de la evaluación y la definición operativa de la variable que interesa medir se debe llevar a cabo determinadas especificaciones relacionadas con el instrumento de medida. En esta fase se debe describir de forma detallada y precisa aspectos concernientes a los requerimientos de administración del instrumento de medida, el tipo, número, longitud, contenido y distribución de los ítems, especificaciones e instrucciones en la entrega del material y aspectos relacionados con la seguridad del mismo.

Los requerimientos de administración del instrumento de medida se refieren a cuál va a ser el soporte de administración (papel o informático), a qué tipo de aplicación se va a realizar (individual o colectiva), y cuándo y en qué lugar se va a administrar el instrumento de medida. Igualmente, se deben especificar los requerimientos cognitivos, de vocabulario y de accesibilidad de los participantes. Es importante llevar a cabo adaptaciones de acceso en aquellos participantes que no puedan desempeñar la tarea en igualdad de condiciones que el resto, por ejemplo disponer de una versión en *Braille* para una persona con deficiencia visual. Todo sistema universitario que se precie debería evaluar en las mismas condiciones y con la misma calidad a todos sus integrantes independientemente de su condición.

En relación con los ítems se debe especificar el tipo, el número, la longitud, el contenido y el orden (disposición) de los mismos, así como el formato de respuesta o el tipo de alternativas que se van a utilizar. Con respecto a este tema, no existen normas universales, todo dependerá de las circunstancias de aplicación, del propósito del constructor y de otras variables.

3.4. Construcción de los ítems

La construcción de los ítems constituye una de las etapas más cruciales dentro del proceso de construcción del instrumento de medida (Downing, 2006; Schmeiser y Welch, 2006). Los ítems son la materia prima, los ladrillos, a partir de la cual se forma un instrumento de evaluación, por lo que una construcción deficiente de los mismos, como no puede ser de otro modo, incidirá en las propiedades métricas finales del instrumento de medida y en las inferencias que se extraigan a partir de las

puntuaciones (Muñiz et al., 2005). Los principios básicos que deben regir la construcción de cualquier banco de ítems son: representatividad, relevancia, diversidad, claridad, sencillez y comprensibilidad (Muñiz et al., 2005). Todos los dominios de la variable de interés deben de estar igualmente representados, aproximadamente con el mismo número de ítems, a excepción de que se haya considerando un dominio más relevante dentro de la variable, y que por lo tanto, deba tener un mayor número de ítems, esto es, una mayor representación. Un muestreo erróneo del dominio objeto de evaluación sería una clara limitación a las inferencias que con posterioridad se dibujen a partir de los datos. Los ítems deben de ser heterogéneos y variados para así recoger una mayor variabilidad y representatividad de la variable de medida. Debe primar la claridad y la sencillez, se deben evitar tecnicismos, dobles negaciones, o enunciados excesivamente prolijos o ambiguos (Muñiz et al., 2005). Del mismo modo, los ítems deben ser comprensibles para la población a la cual va dirigido el instrumento de medida, evitándose en todo momento un lenguaje ofensivo y/o discriminatorio. Ítems con una redacción defectuosa o excesivamente vagos van a incrementar el porcentaje de varianza explicada debido a factores espurios o irrelevantes, con la consiguiente merma de validez de la prueba.

Si los ítems provienen de otro instrumento ya existente en otro idioma y cultura, deberán seguirse las directrices internacionales para la traducción y adaptación de tests (Balluerka, Gorostiaga, Alonso-Arbiol y Haranburu, 2007; Hambleton, Merenda y Spielberger, 2005; Muñiz y Bartram, 2007). En el caso de ítems originales han de seguirse las directrices elaboradas para el desarrollo de ítems de elección múltiple (Downing y Haladyna, 2006; Haladyna, 2004; Haladyna et al., 2002; Moreno et al., 2006; Moreno et al., 2004; Muñiz et al., 2005).

Durante las fases iniciales de la construcción del banco de ítems se recomienda que el número de ítems inicial sea como mínimo el doble del que finalmente se considera que podrían formar parte de la versión final del instrumento de medida. La razón es bien sencilla, muchos de ellos por motivos diferentes (métricos, comprensibilidad, dificultad, etc.) se acabarán desechando, por lo que sólo quedarán aquellos que ofrezcan mejores indicadores o garantías técnicas (sustantivas y métricas). Finalmente, para garantizar la validez de contenido de los ítems (Sireci, 1998) se ha de recurrir a la consulta de expertos y a la revisión exhaustiva de las fuentes bibliográficas, así como a otros instrumentos similares ya existentes. En relación con la valoración de los ítems por parte de los expertos y con la finalidad de una evaluación más precisa y objetiva del conjunto inicial de ítems, se puede pedir a los expertos que juzguen, a partir de un cuestionario, si los ítems están bien redactados para la población de interés, si son o no pertinentes para evaluar una faceta o dominio determinado y si cada ítem representa de forma adecuada la variable o dimensión de interés.

3.5. Producción, base de datos, normas de puntuación y corrección

En esta fase se compone, se edita y se lleva a imprimir la primera versión del instrumento de medida, además de construir la base de datos con la claves de corrección. Este paso ha sido con frecuencia injustamente infraestimado y olvidado, sin embargo es clave, pues el continente bien podría echar a perder el contenido. Buenos ítems pobremente editados dan como resultado un mal test, igual que las malas barricas pueden echar a perder los buenos caldos. Podemos haber construido un buen banco de ítems que de nada servirá si luego éstos se presentan de forma desorganizada, con errores tipográficos, o en un cuadernillo defectuoso. Uno de los errores más frecuentes entre los constructores de tests aficionados es utilizar fotocopias malamente grapadas, con la excusa de que sólo se trata de una versión experimental de la prueba, olvidándose que para las personas que la responden no existen pruebas experimentales, todas son definitivas. El aspecto físico de la prueba forma parte de la validez aparente. Es importante que el instrumento dé la impresión de medir de manera objetiva, rigurosa, fiable y válida la variable de interés. Por otra parte, en esta fase también se debe construir, si fuera el caso, la base de datos donde posteriormente se van a tabular las puntuaciones y a realizar los análisis estadísticos pertinentes así como las normas de corrección y puntuación, por ejemplo si existen ítems que se deben recodificar, si se va a crear una puntuación total o varias puntaciones, etc.

3.6. Estudio piloto cualitativo y cuantitativo

La finalidad de cualquier estudio piloto es examinar el funcionamiento general del instrumento de medida en una muestra de participantes con características semejantes a la población objeto de interés. Esta fase es de suma importancia ya que permite detectar, evitar y

corregir posibles errores así como llevar a cabo una primera comprobación del funcionamiento del instrumento de evaluación en el contexto aplicado. El estudio piloto podría verse como una representación en miniatura de lo que posteriormente va a ser el estudio de campo.

Existen dos tipos de estudio piloto: cualitativo y cuantitativo (Wilson, 2005). El estudio piloto cualitativo permite, a partir de grupos de discusión, debatir en voz alta diferentes aspectos relacionados con el instrumento de medida (p. ej., la detección de errores semánticos, gramaticales, el grado de comprensibilidad de los ítems, las posibles incongruencias semánticas, etc.). Los participantes en este pilotaje pueden ser o no similares a la población objeto de medición. Por su parte, el estudio piloto cuantitativo permite examinar las propiedades métricas del instrumento de medida. En ambos casos se deben anotar de forma detallada todas las posibles incidencias acaecidas durante la aplicación (p. ej., preguntas o sugerencias de los participantes, grado de comprensión de los ítems así como posibles errores o problemas detectados en el instrumento).

A continuación, una vez tabulados los datos, se procede a los análisis de la calidad métrica de los ítems. En función de criterios sustantivos y estadísticos algunos ítems son descartados mientras que otros son modificados. Es importante que el constructor del instrumento de evaluación deje constancia de qué ítems fueron eliminados o modificados y por qué, además de explicitar con claridad el criterio (cualitativo o cuantitativo) por el cual se eliminaron. En este paso si se considera conveniente se pueden incorporar nuevos ítems. Todas las actividades deben ir destinadas a seleccionar los ítems con mayores garantías métricas que maximicen las propiedades finales del instrumento de evaluación. Finalmente, se debe construir una nueva versión del instrumento de medida que es revisada de nuevo por el grupo de expertos y que será la que en última instancia se administre en el estudio final de campo.

3.7. Selección de otros instrumentos de medida convergentes

La selección adecuada de otros instrumentos de evaluación permite recoger evidencias a favor de la validez de las puntuaciones de los participantes (Elosúa, 2003). Es interesante que no se pierda el norte, la finalidad última de todo proceso de construcción de instrumentos de evaluación es siempre obtener mayores evidencias de validez. La selección adecuada de otras variables de interés permite aglutinar diferentes tipos de evidencias que conduzcan a una mejor interpretación de las puntuaciones en el instrumento de medida dentro de un contexto y uso particular. En este sentido, se pueden establecer relaciones con un criterio externo, con otros instrumentos de medida que pretendan medir la misma variable u otras diferentes (lo que anteriormente se había definido como definición sintáctica).

La utilización de materiales complementarios se encuentra claramente influenciada por cuestiones pragmáticas. Una vez más se vuelve a imponer la realidad. La decisión de qué instrumentos se deben utilizar complementariamente con el nuestro está influenciada por las exigencias referidas al tiempo y al lugar. Evidentemente las exigencias de tiempo y las razones éticas no permiten administrar todos los instrumentos que quisiéramos, si bien aquí no se trata de pasar cuantos más mejor sino de seleccionar aquellos de mayor calidad científica, a partir de los cuales se pueda profundizar en el significado de nuestras puntuaciones. Algunos recomendaciones prácticas en la selección de otros instrumentos de medida son: a) que se encuentren validados para población objeto de interés y se conozcan sus propiedades métricas; b) que sean sencillos y de rápida administración y que conlleven un ahorro de tiempo; c) que tengan "coherencia" sustantiva de cara a establecer relaciones entre las variables.

3.8. Estudio de campo

En la fase del estudio de campo se incluye la selección de la muestra (tipo, tamaño y procedimiento), la administración del instrumento de medida a los participantes y el control de calidad y seguridad de la base de datos.

La representatividad y generalizabilidad de nuestros resultados depende en gran medida de que la muestra elegida sea realmente representativa de la población objetivo de estudio. Elegir una muestra pertinente en cuanto a representatividad y tamaño es esencial, si se falla en esto todo lo demás va a quedar invalidado. El muestreo probabilístico siempre es preferible al no probabilístico, para la estimación del tamaño muestral requerido para un determinado error de medida ha de acudirse a los textos especializados, o consultar los expertos en la tecnología de muestreo. Es recomendable que por

cada ítem administrado tengamos al menos 5 ó 10 personas, si bien determinadas técnicas estadísticas pueden reclamar incluso más de cara a una buena estimación de los parámetros.

Las actividades relacionadas con la administración y el uso del instrumento de medida son cruciales durante el proceso de validación (Muñiz y Bartram, 2007; Muñiz et al., 2005). Cuando administramos cualquier instrumento de medida hay que cuidarse de que las condiciones físicas de la aplicación sean las adecuadas (luz, temperatura, ruido, comodidad de los asientos, etc.). Igualmente, las personas encargadas de la administración del instrumento de medida deben establecer una buena relación (*rapport*) con los participantes, estar familiarizados con la administración de este tipo de herramientas, dar las instrucciones a los participantes correctamente, ejemplificar con claridad como se resuelven las preguntas, supervisar la administración y minimizar al máximo las posibles fuentes de error. Por todo ello es recomendable elaborar unas pautas o directrices que permitan estandarizar la administración del instrumento de medida.

El control de calidad de la base de datos es otro tema a veces poco valorado en el proceso de construcción de instrumentos de medida. Por control de calidad nos referimos a una actividad que tiene como intención comprobar que los datos introducidos en la base de datos se correspondan, de hecho, con las puntuaciones de los participantes en la prueba. Frecuentemente cuando introducimos las puntuaciones de los participantes en una base de datos se pueden cometer multitud de errores, por ello es altamente recomendable comprobar de forma rigurosa que los datos se han introducido correctamente. Una estrategia sencilla a posteriori que se puede utilizar es extraer al azar un cierto porcentaje de los participantes y comprobar la correspondencia entre las puntuaciones en la prueba y la base de datos. No obstante los mejores errores son los que no se cometen, así que hay que poner todos los medios para minimizar los errores a la hora de construir la base de datos.

3.9. Estimación de las propiedades métricas

Una vez administrado el instrumento de medida a la muestra de interés se procede al estudio de las propiedades métricas del mismo: análisis de los ítems, estudio de la dimensionalidad, estimación de la fiabilidad, obtención de evidencias de validez y construcción de baremos.

En esta fase debe primar por encima de todo el rigor metodológico. Todos los pasos y decisiones que se tomen se deben describir con claridad y deben de estar correctamente razonadas. En un primer lugar, se deben analizar los ítems tanto a nivel cualitativo como a nivel cuantitativo. Para seleccionar los mejores ítems desde el punto de vista métrico se pueden tener en cuenta el índice de dificultad (cuando proceda), el índice de discriminación, las cargas factoriales y/o el funcionamiento diferencial de los ítems (Muñiz et al., 2005). No se debe perder de vista que la finalidad del análisis métrico de los ítems no debe ser otro que maximizar o potenciar las propiedades métricas del instrumento de medida; no obstante no existen reglas universales y las consideraciones estadísticas no garantizan unos resultados con significación conceptual, por lo que uno debería tener presente también los aspectos sustantivos (Muñiz et al., 2005). Una vez seleccionados los ítems, se procede al estudio de la dimensionalidad del instrumento para conocer su estructura interna. 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. El análisis factorial y el análisis de componentes principales son las técnicas más utilizadas para examinar la estructura interna, si bien no son las únicas (Cuesta, 1996). Una vez determinada la dimensionalidad del instrumento de medida se lleva a cabo una estimación de la fiabilidad, para lo cual se pueden seguir diversas estrategias, tanto desde el punto de vista de la teoría clásica de los tests como de la teoría de respuesta a los ítems (Muñiz, 1997, 2000). Posteriormente, y de cara a obtener evidencias de validez, se debe observar la relación del instrumento de medida con otros instrumentos de evaluación, y finalmente, se lleva a cabo una baremación del instrumento de medida donde se establecen puntos de corte normativos. Los desarrollos estadísticos y técnicos en este campo son notables, incorporándose cada vez más a menudo los métodos estadísticos robustos (Erceg-Hurn y Mirosevich, 2008), el análisis factorial confirmatorio (Brown, 2006; Kline, 2005) y el funcionamiento diferencial de los ítems, por citar sólo tres casos (Muñiz et al., 2005).

3.10. Versión definitiva, informe final y manual del instrumento de medida

En último lugar, se procede a la elaboración de la versión definitiva del instrumento de medida, se envía un informe de resultados a las partes interesadas (alumnos, profesores,

departamentos, etc.) y se elabora el manual del mismo que permita su utilización a otras personas o instituciones interesadas. El manual de la prueba debe de recoger con todo detalle todas las características relevantes de la prueba. Como se comentó anteriormente, todo proceso de evaluación es necesario que concluya en un *feedback* (rápido, conciso y claro) a las partes implicadas y con una propuesta de planes de mejora. Finalmente y aunque sea la última fase, esto no quiere decir que el proceso de validación concluya aquí, posteriores estudios deberán seguir recogiendo evidencias de validez que permitan tomar decisiones fundadas a partir de las puntuaciones de los individuos.

4. A MODO DE CONCLUSIÓN

En las líneas precedentes ya se ha indicado que el proceso de evaluación universitaria es complejo, dada la cantidad de aspectos diferentes susceptibles de ser evaluados. Para llevar a cabo una evaluación universitaria rigurosa no sólo es necesario, que lo es, disponer de instrumentos de evaluación técnicamente solventes, además hay que manejar un modelo de evaluación integral que dé una respuesta a los siguientes interrogantes: 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 qué opinión tienen las partes implicadas sobre la evaluación. Tras un breve repaso por esos siete aspectos nos hemos centrado en la descripción de los diez pasos básicos que habría que seguir para desarrollar un instrumento de medida objetivo y riguroso. Estos pasos no se pueden abordar en profundidad desde un punto de vista técnico en un breve artículo como éste, no se trata de eso, sino de poner a disposición de los gestores y profesionales una guía general que les permita obtener una visión panorámica de las actividades implicadas en el desarrollo de los instrumentos de medida. Esperamos haber sido capaces de transmitir la idea de que el campo de la elaboración de instrumentos de medida en el contexto de la Evaluación Universitaria está altamente desarrollado y es necesario acudir a personal cualificado para su desarrollo adecuado, constituyendo una temeridad dejarlo en manos de aficionados bienintencionados.

5. BIBLIOGRAFÍA

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*, 153-166.

American Educational Research Association, American Psychological Association, y National Council on Measurement in Education (1999). *Standars for Educational and Psychological Testing*. Washington, DC: Author.

Anastasi, A., y Urbina, S. (1998). *Los tests psicológicos*. México: Prentice Hall.

Balluerka, N., Gorostiaga, A., Alonso-Arbiol, I., y Haranburu, M. (2007). La adaptación de instrumentos de medida de unas culturas a otras: una perspectiva práctica. *Psicothema*, 124-133.

Beran, T., y Violato, C. (2005). Rating of university teacher instruction: How much do student and course characteristics really matter? *Assessment & Evaluation in Higher Education, 30*, 593-601.

Brennan, R. L. (2006). *Educational Measurement*. Washington, DC: American Council on Education/Praeger.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Buela-Casal, G., Bermúdez, M. P., Sierra, J. C., Quevedo-Blasco, R., y Castro-Vázquez, A. (2009). Ranking de 2008 en productividad en investigación de las universidades públicas españolas. *Psicothema, 21*.

Centra, J. A. (1993). *Reflective faculty evaluation*. San Francisco: Jossey-Bass.

Carretero-Dios, H., y Pérez, C. (2005). Normas para el desarrollo y revisión de estudios instrumentales. *International Journal of Clinical and Health Psychology, 5*, 521-551.

Clark, L. A., y Watson, D. (1995). Constructing Validity: Basic issues in objective scale development. *Psychological Assessment 7*, 309-319.

Cuesta, M. (1996). Unidimensionalidad. En J. Muñiz (Ed.), *Psicometría*. Madrid: Universitas. (pags. 239-292).

Dolores, M., y Padilla, J. L. (2004). Técnicas psicométricas: los tests. En R. Fernández-Ballesteros (Ed.), *Evaluación psicológica:*

Conceptos, métodos y estudio de casos (pp. 323-355). Madrid: Pirámide.

Downing, S. M. (2006). Twelve steps for effective test development. En S. M. Downing y T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum Associates.

Downing, S. M., y Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Elosúa, P. (2003). Sobre la validez de los tests. *Psicothema, 15*, 315-321.

Erceg-Hurn, D. M., y Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63*, 591-601.

Fernández, J. (2008). *Valoración de la calidad docente*. Madrid: Editorial Complutense.

Fernández, J., Mateo, M. A., y Muñiz, J. (1995). Evaluation of the academic setting in Spain. *European Journal of Psychological Assessment*(11), 133-137.

Fernández, J., Mateo, M. A., y Muñiz, J. (1996). Valoración por parte del profesorado de la evaluación docente realizada por los alumnos. *Psicothema, 8*, 167-172.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test item (3ª ed.)*. Hillsdale, NJ: LEA.

Haladyna, T. M., Downing, S. M., y Rodríguez, M. C. (2002). A review of multiple-choice item-writing guidelines. *Applied Measurement in Education, 15*(3), 309-334.

Hambleton, R. K., Merenda, P. F., y Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment* London: Lawrence Erlbaum Associates

Institute of Higher Education, Shangai Jiao Tong University (2008). Academic Ranking of World Universities. *http://ed.sjtu.edu.cn/rank/2008/2008Main.htm*

International Ranking Expert Group (2006). Berlin Principles on Ranking of Higher Education Institutions. *http://www.che.de/downloads/Berlin_Principles_IREG_534.pdf*

Joint Committee on Standards for Educational Evaluation. (2003). *The student evaluation standards*. Thousand Oaks, CA: Corwin Press.

Kane, M. T. (2006). Validation. En R. L. Brennan (Ed.), *Educational measurement (4th ed.)* (pp. 17-64). Westport, CT: American Council on Education/Praeger.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2 ed.). New York: The Guilford Press.

Lord, F. M., y Novick, M. R. (1968). *Statistical theories of mental test scores*. New York: Addison-Wesley.

Marsh, H. W., y Roche, L. A. (2000). Effects of grading leniency and low workloads on students' evaluations of teaching: Popular myths, bias, validity or innocent bystanders. *Journal of Educational Psychology, 92*, 202-228.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research 45* 35-44.

Morales, P., Urosa, B., y Blanco, A. B. (2003). *Construcción de escalas de actitudes tipo Likert*. Madrid: La Muralla.

Moreno, R., Martínez, R., y Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology, 2*, 65-72.

Moreno, R., Martínez, R. J., y Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema, 16*(3), 490-497.

Muñiz, J. (Ed.) (1996). *Psicometría*. Madrid: Universitas.

Muñiz, J. (1997a) Introducción a la teoría de respuesta a los ítems. Madrid: Pirámide.

Muñiz, J. (1997b). Aspectos éticos y deontológicos de la evaluación psicológica. En A. Cordero (ed.), *La evaluación psicológica en el año 2000*. Madrid: Tea Ediciones.

Muñiz, J. (2000). *Teoría Clásica de los Tests*. Madrid: Pirámide.

Muñiz, J. (2004). La validación de los tests. *Metodología de las Ciencias del Comportamiento, 5*, 121-141.

Muñiz, J., y Bartram, D. (2007). Improving international tests and testing. *European Psychologist, 12*, 206-219.

Muñiz, J., Fidalgo, A. M., García-Cueto, E., Martínez, R., y Moreno, R. (2005). *Análisis de los ítems*. Madrid: La Muralla.

Nunnally, J. C., y Bernstein, I. J. (1995). *Teoría psicométrica*. México: McGraw Hill.

Padilla, J. L., Gómez, J., Hidalgo, M. D., y Muñiz, J. (2006). La evaluación de las consecuencias del uso de los tests en la teoría de la validez. *Psicothema, 19*, 307-312.

Padilla, J. L., Gómez, J., Hidalgo, M. D., y Muñiz, J. (2007). Esquema conceptual y procedimientos para analizar la validez de las consecuencias del uso de los test. *Psicothema, 19*, 173-178

Schmeiser, C. B., y Welch, C. (2006). Test development. En R. L. Brennan (Ed.), *Educational Measurement (4th ed.)* (pp. 307-353). Westport, CT: American Council on Education/Praeger.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5*, 299-321.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher 36*, 477-481.

Sireci, S. G., y Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity *Educational Measurement: Issues and Practice 25*, 27-34.

Smith, G. T., Fischer, S., y Fister, S. M. (2003). Incremental validity principles in test construction. *Psychological Assessment, 15*, 467-477.

Smith, S. T. (2005). On construct validity: Issues of method measurement. *Psychological Assessment, 17*, 396-408.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. En C. R. Rao y S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 45-79). Amsterdam, Netherlands: Elsevier Science.

Nota. Este trabajo ha sido financiado por el Ministerio Español de Ciencia e Innovación, referencia PSI2008-03934 y CIBERSAM Universidad de Oviedo.