Papeles del Psicólogo, 2013. Vol. 34(2), pp. 82-90
http://www.papelesdelpsicologo.es

SEGUNDA EVALUACIÓN DE TESTS EDITADOS EN ESPAÑA

SECOND EVALUATION OF TESTS PUBLISHED IN SPAIN

Vicente Ponsoda[1] y Pedro Hontangas[2]
[1]*Universidad Autónoma de Madrid.* [2]*Universidad de Valencia*

El artículo describe los resultados de la segunda evaluación de tests psicológicos editados en España. La Comisión de Tests del Colegio Oficial de Psicólogos decidió que se evaluasen 12 tests, seleccionados principalmente por su novedad y amplio uso. Cada test ha sido evaluado por dos expertos. Al igual que en la primera evaluación (Muñiz, Fernández-Hermida, Fonseca-Pedrero, Campillo-Álvarez y Peña-Suárez, 2011), los evaluadores hacían su trabajo respondiendo a las preguntas del Cuestionario para la Evaluación de los Tests (Prieto y Muñiz, 2000), que adapta al contexto español el modelo elaborado por la Federación Europea de Asociaciones de Psicólogos Profesionales. De cada test se ha evaluado la calidad de los materiales y documentación, la fiabilidad de sus puntuaciones, la cobertura de los estudios de validación, la calidad de los baremos, etc. Los revisores informaron también acerca de la idoneidad del instrumento y proceso seguido en la evaluación. Se aportan sugerencias que pudieran ser útiles para mejorar las evaluaciones futuras.
Palabras clave: *Tests, Uso de los tests, Evaluación de tests, Psicometría.*

This article describes the results of the second evaluation of psychological tests published in Spain. The Committee on Testing of the Spanish Psychological Association agreed on assessing 12 tests, selected mainly for their novelty and wide use. Each test has been evaluated by two experts. As in the first evaluation (Muñiz, Fernández-Hermida, Fonseca-Pedrero, Campillo-Álvarez and Peña-Suárez, 2011), assessments were made by responding to the Questionnaire for the Assessment of Tests (Prieto and Muñiz, 2000), which adapts to the Spanish context the assessment model developed by the European Federation of Psychologists' Associations. Results are provided in both absolute and relative terms, as they are compared to those of the first evaluation. They refer to the quality of documentation and materials, the coverage of the validation studies, reliability, norms, etc. Reviewers were also asked about the suitability of the instrument and procedure used to conduct the assessment. Suggestions are provided that may be useful to improve next test evaluations.
Key Words: *Tests, Test use, Test evaluation, Psychometrics.*

Los tests psicológicos son herramientas muy utilizadas por la mayoría de los psicólogos. Con ellos toman decisiones de considerable impacto social en los diversos campos de actuación de la Psicología, como el educativo, clínico, social, organizacional y jurídico. No resulta extraño entonces que varios países europeos realicen procesos sistemáticos de evaluación de sus tests. Es el caso del Reino Unido, Alemania, Noruega… y, en especial, Holanda. Evers, Sijtsma, Lucassen y Meijer (2010) informan de las principales características del proceso de evaluación holandés. Hace más de 40 años, la Comisión de Tests del Colegio de Psicólogos holandés puso en marcha una primera evaluación de la calidad de sus tests, dando lugar a un primer libro de evaluaciones, publicado en 1969. A esta primera publicación han seguido cinco más, la última en 2009. Entre 1982 y 2010, el número de tests revisados ha sido 878, que son prácticamente todos los editados. Por tanto, en Holanda, quien esté interesado en aplicar un test puede encontrar una evaluación independiente y rigurosa de su calidad y propiedades.

Muñiz y Fernández-Hermida (2010) muestran que la opinión de los psicólogos españoles sobre el uso de los tests es claramente positiva. En una escala de 1 ("desacuerdo total") a 5 ("totalmente de acuerdo"), fue 4.41 la media en el ítem "Utilizados correctamente los tests son de gran ayuda para el psicólogo". También están de acuerdo (media = 4.13) con el ítem "El Colegio Oficial de Psicólogos debería de ejercer un papel más activo para regular y mejorar el uso que se hace de los tests". En el mismo estudio, los acuerdos anteriores se tornan en ligero desacuerdo (media = 2.71) cuando la frase es "Los profesionales disponen de suficiente información (revisiones independientes, investigaciones, documentación…) sobre la calidad de los tests editados en nuestro país". Estos resultados animaron al Colegio Oficial de Psicólogos (COP), por medio de su Comisión de tests

Correspondencia: Vicente Ponsoda. *Facultad de Psicología. Universidad Autónoma de Madrid. C/ Iván Pavlov 6. 28049 Madrid. España. E-mail: Vicente.ponsoda@uam.es*

(CT), a empezar la revisión sistemática de los tests editados en España, siguiendo la estela de las evaluaciones que están haciendo otros países. En 2010 se puso en marcha el proceso y se revisaron 10 tests. Los principales resultados se han publicado en esta misma revista (Muñiz, Fernández-Hermida, Fonseca-Pedrero, Campillo-Álvarez y Peña-Suárez, 2011). En la web del COP (http://cop.es/8) están disponibles las revisiones completas de cada test. La experiencia fue valorada positivamente por la CT y se decidió realizar una segunda evaluación. El presente trabajo describe sus principales resultados.

PROCESO DE EVALUACIÓN

El proceso seguido en esta segunda evaluación coincide básicamente con el de la primera, con alguna diferencia que se indicará oportunamente. En primer lugar, la CT decidió qué tests debían ser revisados. Decidió que fuesen 12. Once son pruebas comercializadas por los editores de tests que integran la CT (3 de EOS, 3 de Pearson y 5 de TEA). La prueba duodécima es la escala EPV-R (Echeburúa, Amor, Loinaz y Corral, 2010), no comercializada, que viene siendo aplicada para la detección de riesgo alto de violencia grave contra la pareja en muchas comisarías de policía. La lista de los 12 tests se muestra en la Tabla 1. El siguiente paso fue decidir quién podría ser el coordinador. A propuesta de la CT, el primer firmante de este trabajo aceptó serlo.

A continuación comenzó la búsqueda de los revisores. Como en la primera evaluación, se pensó que lo apropiado era que un evaluador fuese más experto en cuestiones psicométricas y el otro lo fuese más en las variables sustantivas que el test medía. Se puso especial cuidado en los siguientes tres asuntos: En primer lugar, se procuró que el experto en Psicometría evaluase el test que por contenido pudiese interesarle más y conocer mejor. Algo similar se hizo con los expertos en el contenido. Del conjunto de expertos en un contenido particular fue elegido el que tuviese publicaciones más relacionadas con los tests. En segundo lugar, el coordinador decidió no recurrir a ninguno de los revisores participantes en la primera evaluación, con la idea de ir poco a poco generando un banco de revisores; pues la CT pretende continuar con el proceso de evaluación de tests en años sucesivos. Por último, se procuró que los revisores no tuviesen relación directa con los autores de los tests. De hecho, en la carta de invitación se les indicaba que no aceptasen participar en la revisión si dudaban de que pudiesen hacer una valoración objetiva.

Una primera lista con los 2 posibles revisores de cada test fue analizada por otros 2 miembros de la CT. Advirtieron que un revisor no era apropiado y la lista fue modificada. A continuación, el coordinador les invitó a participar. De las 24 invitaciones, el número de rechazos fue 2. En un caso, por razones personales muy compren-

TABLA 1
RELACIÓN DE TESTS EVALUADOS

BAI	Inventario de ansiedad de Beck
BAS-II	Escalas de aptitudes intelectuales
BDI-II	Inventario de depresión de Beck-II
CEAM	Cuestionario de estrategias de aprendizaje y motivación
CompeTEA	Cuestionario para la evaluación de las competencias en el contexto laboral
EPV-R	Escala de predicción del riesgo de violencia grave contra la pareja - Revisada
ESCOLA	Escala de conciencia lectora
ESPERI	Cuestionario para la detección de los trastornos del comportamiento en niños y adolescentes
Merrill-Pamer-R	Escalas de desarrollo Merrill-Palmer revisadas
PAI	Inventario de evaluación de la personalidad
RIAS/RIST	Escalas de inteligencia de Reynolds/Test de inteligencia breve de Reynolds
WNV	Escala no verbal de aptitud intelectual de Wechsler

TABLA 2
REVISORES QUE LLEVARON A CABO LA EVALUACIÓN DE LOS TESTS

Revisor	Afiliación
Francisco José Abad García	Universidad Autónoma de Madrid
Jesús Alonso Tapia	Universidad Autónoma de Madrid
Jesús Alvarado Izquierdo	Universidad Complutense de Madrid
Juan Antonio Amador Campos	Universidad de Barcelona
Ramón Arce Fernández	Universidad de Santiago de Compostela
Roberto Colom Marañón	Universidad Autónoma de Madrid
Pere Joan Ferrando Piera	Universidad Rovira y Virgili
Eduardo García Cueto	Universidad de Oviedo
Paula Elosua Oliden	Universidad del País Vasco
Pedro Hontangas Beltrán	Universidad de Valencia
Fernando Jiménez Gómez	Universidad de Salamanca
Antonio Lobo Satué	Universidad de Zaragoza
Raúl López Antón	Universidad de Zaragoza
Ramón López Sánchez	Universidad Complutense de Madrid
Antonio Maldonado Rico	Universidad Autónoma de Madrid
María Rosario Martínez Arias	Universidad Complutense de Madrid
Julio Olea Díaz	Universidad Autónoma de Madrid
José Olivares Rodríguez	Universidad de Murcia
José Luis Padilla García	Universidad de Granada
Lilisbeth Perestelo Pérez	Servicio Canario de la Salud
Gerardo Prieto Adánez	Universidad de Salamanca
María Ángeles Quiroga Estévez	Universidad Complutense de Madrid
Jordi Renom Pinsach	Universidad de Barcelona
Jesús Salgado Velo	Universidad de Santiago de Compostela
Carme Viladrich Segués	Universidad Autónoma de Barcelona

sibles, y en el segundo por considerar el evaluador que su evaluación podría no ser objetiva. La selección de revisores, que es un punto trascendental, es uno de los aspectos que ha resultado más satisfactorio. Como se puede ver en la Tabla 2, su calidad científica es difícilmente mejorable y así lo ha sido también su involucración y buen hacer en las distintas fases del proceso. El coordinador aprovecha estas líneas para expresar su profundo agradecimiento a todos ellos.

Los editores pusieron a disposición del COP dos ejemplares completos de cada test. El COP envió uno a cada revisor. En relación al test EPV-R, no comercializado, el coordinador explicó al autor principal que su test iba a ser revisado y le pidió que indicase la documentación mínima sobre la que los revisores pudieran hacer su trabajo. Su respuesta fue positiva y muy colaboradora, y así se determinó la documentación que los dos revisores habrían de evaluar.

Como en la primera evaluación, la recogida de las evaluaciones se ha de hacer respondiendo al Cuestionario para la Evaluación de los Tests (CET), que adapta al contexto español el modelo de evaluación de tests desarrollado por la Federación Europea de Asociaciones de Psicólogos Profesionales (Prieto y Muñiz, 2000).

En las siguientes líneas se ofrece una breve descripción del CET, pues el resto del trabajo hace continuas referencias a sus características. Tiene tres apartados. El primero (Descripción general) consta de 31 preguntas sobre el nombre del test, autores, fecha de publicación, descripción de qué mide, áreas de aplicación, formato de los ítems, modos de administración, precio, etc. El segundo apartado (Valoración de las características) consta, a su vez, de varios sub-apartados y contiene 35 preguntas en total. Las preguntas del primer sub-apartado tienen que ver con la calidad de los materiales, documentación e instrucciones, la fundamentación teórica, la facilidad de administración, la calidad del proceso de adaptación y el análisis formal y psicométrico de los ítems. Las preguntas del segundo sub-apartado (validez) evalúan la validez de contenido, de constructo, predictiva y el sesgo de los ítems. El tercer sub-apartado (fiabilidad) incluye preguntas sobre indicadores de equivalencia, consistencia interna y estabilidad. El último sub-apartado evalúa la calidad de los baremos. En el tercer y último apartado (Valoración global) el revisor ha de aportar su evaluación global, la relación de puntos fuertes y débiles del test y ha de rellenar dos tablas resumen. La primera ha de completarse a partir de las 31 preguntas del apartado Descripción general y la segunda requiere la evaluación en 12 características (listadas en la primera columna de la Tabla 3) que resumen las evaluaciones proporcionadas en el apartado Valoración de las características. De las 68 preguntas del CET, 25 son cuantitativas y se han de responder en una escala tipo Likert de 5 categorías (desde 1, "inadecuada", a 5, "excelente"). En cada pregunta se detalla el significado concreto de "excelente". El resto son preguntas abiertas.

Se comunicó a cada revisor que su tarea principal era aplicar el CET al test que se le había asignado y se le informó sucintamente también de las siguientes fases del proceso. Recibidas las evaluaciones, fue tarea del coordinador generar una evaluación conjunta a partir de las dos recibidas.

El penúltimo paso fue enviar las evaluaciones conjuntas a los editores y autores del test no comercializado para recabar su parecer y comentarios. Las respuestas han variado apreciablemente en extensión y en acuerdo con lo indicado en la evaluación enviada. En nuestra opinión la participación del editor/autor es un elemento muy importante para asegurar la calidad de la revisión final. A partir de las respuestas recibidas, el coordinador modificó las evaluaciones cuando lo consideró oportuno y generó las evaluaciones finales. El último paso fue presentarlas a la CT, para que las conociese y diese el visto bueno, antes de hacerlas públicas.

RESULTADOS

Como se ha indicado, el CET requiere evaluaciones cualitativas y cuantitativas. No todas las preguntas pueden puntuarse. Por ejemplo, una pregunta requiere que se evalúe la calidad de la adaptación, cuando no todos los tests evaluados son adaptaciones. En otras, el manual puede no ofrecer información sobre lo que se pregunta. Para obtener alguna evidencia de la fiabilidad entre los evaluadores, se ha calculado la correlación entre las puntuaciones otorgadas por los 2 evaluadores de cada test en las preguntas (5) en las que todos los tests evaluados tienen puntuaciones. La mediana de las 12 correlaciones, tantas como tests evaluados, es 0.61.

Los principales resultados se muestran en la Tabla 3, que contiene las evaluaciones de los 12 tests en cada una de las 12 características incluidas en la tabla resumen de valoración. La penúltima columna contiene las medias de las evaluaciones en cada característica (su valor mínimo es 1 y el máximo, 5). La mayor media es 5, y corresponde a la característica "Análisis del sesgo",

si bien se ha obtenido sólo a partir de un test. En la primera evaluación, ninguno de los 10 tests fue evaluado en esta característica. Las siguientes mejores medias son 4.32 y 4.29 y corresponden a "Calidad de los materiales y documentación" y "Adaptación española". La menor es 3.40 y corresponde a la característica "Fiabilidad que informa de la estabilidad". El punto neutro de la escala de respuesta es 3 ("Adecuada"). Incluso la menor media supera este punto neutro. Las dos siguientes peores son 3.50 y 3.55 que corresponden a "Validez de contenido" y "Validez predictiva", respectivamente.

En la primera evaluación las características que resultaron mejor evaluadas, con medias 4.5 y 4.35, fueron la "Fiabilidad mediante indicadores de equivalencia" y la "Calidad de los materiales y documentación", respectivamente. La que resultó peor, con media 3.5, fue el indicador de "Fiabilidad que informa de la estabilidad", como ha ocurrido en la segunda evaluación. Comparando los resultados de las 2 últimas columnas se aprecia que en "Validez de contenido" ha resultado claramente peor la segunda evaluación, mientras en "Validez de constructo" ha resultado peor la primera.

La media total de la segunda evaluación ha sido 4.02, ligeramente por encima de la media de la primera evaluación (3.96). Eliminando de la segunda evaluación la característica "Análisis del sesgo", que solo ha sido evaluada en uno de los 22 tests de ambas evaluaciones, la media de la segunda evaluación quedaría en 3.93. La evaluación media queda muy cerca de 4, valor al que corresponde la etiqueta "Buena" en la escala de respuesta.

Evers et al. (2010) informan de las medias de las evaluaciones realizadas por los revisores holandeses. La escala de respuestas en este caso es de 3 categorías: 1 ("Insuficiente"), 2 ("Suficiente") y 3 ("Bueno"). La media de los últimos 540 tests revisados es 2.03, muy próxima al punto neutro de la escala de respuestas (2), cuando las medias españolas están claramente por encima del punto neutro de la correspondiente escala. Son varias las diferencias entre el sistema holandés de revisión y el español, por lo que no resulta fácil explicar las diferencias encontradas. Una importante es que los holandeses evalúan en 7 criterios, mientras que las evaluaciones españolas mostradas en la Tabla 3 lo hacen en 12, lo que cuestiona que el significado de las puntuaciones sea realmente el mismo. En la Tabla 3, se ha utilizado un guión para indicar ausencia de información o no procede. Cabe la duda de si algunos de los guiones que indican ausencia de información no debieran ser más bien una puntuación baja. De serlo, por supuesto, las medias serían menores.

TABLA 3
RESUMEN DE LAS CALIFICACIONES DE LOS TESTS AVALUADOS

	Tests evaluados												**Media 2012**	Media 2011
Características	BAI	BAS-II	BDI-II	CEAM	Compe TEA	EPV-R	ESCOLA	ESPERI	Merrill Palmer-R	PAI	RIAS RIST	WNV		
Calidad de los materiales y documentación	4.5	4.5	5	3.5	4.5	-	3.5	3	5	4.5	4.5	5	4.32	4.35
Fundamentación teórica	4	5	5	4	4	3	3	3	4	5	4	4	4.00	4.20
Adaptación española	4	5	4	-	-	-	-	-	5	4	4	4	4.29	4.25
Análisis de los ítems	4	5	4	4	4	3	4	3	4	4	4	-	3.91	3.58
Validez de contenido	3	4	4	-	3.5	2.5	3.5	2	4	4	4	4	3.50	4.25
Validez de constructo	4.5	5	5	4	4	2	2.5	4	4.5	4.5	5	5	4.17	3.60
Análisis del sesgo	-	5	-	-	-	-	-	-	-	-	-	-	5.00	-
Validez predictiva	4	4	5	2.5	3.5	3	3.5	-	-	3	4	3	3.55	3.57
Fiabilidad: equivalencia	-	-	-	-	-	-	4	-	-	-	-	-	4.00	4.50
Fiabilidad: consistencia interna	4.5	4	5	4	3	3.5	4	4	4.5	4.5	4.5	4.5	4.17	3.75
Fiabilidad: estabilidad	-	3	-	-	-	-	-	-	4.5	3.5	3	3	3.40	3.50
Baremos	-	4	-	4	4	-	3	4	4	4	4.5	4	3.94	4.00
Media global													4.02	3.96

Notas: Las puntuaciones de la tabla están dadas en una escala cuyos 5 valores son: 1= inadecuado, 2=adecuado pero con algunas carencias, 3= adecuado, 4= bueno, 5= excelente. Cuando aparece un guión (-) significa que no se aporta información o no procede.

COMENTARIOS SOBRE EL CET Y EL PROCESO DE EVALUACIÓN DE TESTS

En relación al CET

Muñiz et al. (2011) indican que convendría modificar el CET a la luz de los cambios introducidos en el nuevo modelo de evaluación europeo (Evers et al., 2010) y mejorar las preguntas que habían resultado en alguna medida problemáticas en la primera evaluación. A partir de los comentarios de los revisores de la segunda evaluación, podemos añadir algunas sugerencias adicionales.

En cada una de las 25 preguntas cuantitativas del CET hemos obtenido la varianza de las 2 evaluaciones dadas por los 2 revisores de cada test. A continuación hemos obtenido la media de las 12 varianzas (una por test) como indicador de la "ambigüedad/dificultad de aplicación" de cada pregunta. Cabe esperar que el acuerdo entre los 2 evaluadores sea menor (y sea mayor la varianza) en las preguntas problemáticas. De las 25 preguntas, las tres que resultaron con mayor varianza media fueron la 2.9.1 ("Análisis de los ítems"), la 2.10.1.2 ("Número de expertos consultados en la validación de contenido") y la 2.10.3.2 ("Tamaño de las muestras en la validación predictiva"). El desacuerdo entre los revisores en la evaluación del análisis de ítems pudiera deberse a que los expertos en Psicometría evalúan específicamente el análisis *psicométrico* de ítems, mientras que los expertos en contenido pudieran haber evaluado la calidad de los ítems sin prestar tanta atención a que el manual proporcione los indicadores de discriminación, de dificultad, detalles de los ítems eliminados, etc. Las preguntas referidas a la validez de contenido han resultado en cierta medida problemáticas; pues, para algunos evaluadores, la información que el manual ofrece, cuando se describe el desarrollo del test, sobre la tabla de especificaciones y los procesos seguidos para comprobar que los ítems se relacionan con el constructo que se pretende medir es evidencia suficiente de validez de contenido. Sin embargo, para otros evaluadores, un estudio de validez de contenido debe ser un estudio posterior al desarrollo del test que pretende mostrar evidencias de si el test evalúa realmente las partes relevantes del constructo de interés. Si existe cierto desacuerdo en qué se entiende por validez de contenido no es de extrañar que también lo exista en el número de expertos involucrados en dicha validación. Algo similar puede decirse de la validación predictiva. La línea de separación entre la validez de constructo convergente y la validez de criterio es a menudo muy delgada. Los manuales presentan a veces en el apartado de validez predictiva estudios que pudieran ser considerados de validación convergente. No resulta entonces raro que los evaluadores muestren desacuerdo a la hora de informar del tamaño de la muestras.

Un asunto al que futuras ediciones del CET convendría que prestasen más atención es el de qué peso dar a los datos originales frente a los obtenidos en la adaptación. El CET incorpora una pregunta para evaluar la calidad de la adaptación. Los manuales suelen ofrecer muchos resultados obtenidos con el test original y generalmente menos con el test adaptado; pues, como cabe esperar, el test original lleva más tiempo disponible y se ha aplicado más veces. Por ejemplo, en el apartado de validez de constructo es frecuente que el manual ofrezca muchos estudios hechos con el test original y algunos hechos en España ¿Qué peso se ha de dar a unos y a otros en la evaluación de la validez de constructo? ¿Hay que tener en cuenta en la evaluación todos los estudios o solo los segundos? Este asunto puede estar detrás de algunas de las discrepancias observadas entre los evaluadores.

En línea con lo sugerido por Muñiz et al. (2011) sobre la primera evaluación, hay que indicar con más claridad cómo debe proceder el evaluador. Tres evaluadores de la segunda evaluación modificaron las opciones del CET cuando no encontraban alguna que se ajustase a lo que querían decir. Estaría bien incluir un conjunto de instrucciones generales, indicando que las opciones del CET no se deben modificar, dando pautas sobre si el evaluador puede o no añadir una nota explicativa o justificativa de las evaluaciones cuantitativas que otorga, que no se deben dejar preguntas sin contestar, etc. Quizás convenga añadir un glosario con la definición de los términos psicométricos que puedan plantear dudas de comprensión. Como Prieto y Muñiz (2000) sugieren, podría contemplarse la posibilidad de una administración informatizada del CET. Haría más uniforme el proceso de respuesta, admitiría solo una o más de una respuesta dependiendo de la pregunta, podría dar el significado de algún término pulsando sobre él, calcularía automáticamente la puntuación en las preguntas en las que se espera que la puntuación sea la media de las puntuaciones asignadas a otras preguntas, etc.

Cabe plantearse si el actual CET resulta adecuado para todos los tipos de tests. Para evaluar las ayudas a la interpretación de las puntuaciones, el CET contiene una sección de baremos, con 4 preguntas. Sin embargo, en

algunas situaciones (escalas clínicas, por ejemplo) tiene más sentido y es más frecuente establecer puntos de corte que permiten clasificar la puntuación obtenida en alguno de los grupos de interés. ¿Sería mejor que el CET incluyera un apartado de interpretación de puntuaciones que permitiera evaluar además otras estrategias de interpretación, alternativas a los baremos? Hay pruebas (Evers et al. 2010) que no están pensadas para predecir resultados externos y en las que tiene poco sentido la validez predictiva. En las escalas clínicas, se suelen ofrecer resultados sobre la capacidad de la prueba para predecir la pertenencia a distintos grupos, que no requieren calcular correlaciones. La pregunta con la que se puntúa la capacidad predictiva del test ("Mediana de las correlaciones del test con los criterios") no parece apropiada en este caso.

Las baterías plantean algunos problemas específicos. El CET indica en una nota lo siguiente: "Si el test está compuesto de subtests heterogéneos en su formato y características, rellena un cuestionario para cada subtest". Conviene destacar que no resulta similar el trabajo que supone al revisor un manual de 50 páginas (BDI-II, por ejemplo) que una batería (BAS-II, por ejemplo, con varios y extensos manuales y distintas pruebas). En algún caso, el revisor dijo al coordinador que si tenía que hacerlo como se indica en la nota, no podría hacer la revisión. Convendría considerar si es adecuado eliminar esa nota e indicar claramente qué información complementaria se ha de aportar en el caso de baterías y en qué preguntas, teniendo presente que sea lo más razonable posible la cantidad de trabajo que se pide al revisor.

Para terminar, a la lista de asuntos que han planteado alguna dificultad en la primera y segunda evaluaciones, añadiríamos algunas sugerencias más: ¿Sería mejor desglosar validez de constructo en estructura interna y relación con otras variables? Algunas pruebas tienen más ítems de los que se han de administrar, pues el aplicador ha de seleccionar los apropiados a cada evaluado. Convendría indicar cómo proceder en ese caso (por ejemplo, indicando el número de ítems disponibles y el máximo posible de ítems a aplicar en cada prueba). También en la segunda evaluación, la pregunta del CET sobre el "procedimiento de corrección" ha planteado alguna dificultad, pues a veces se confunde "automatizada por ordenador" con "efectuado exclusivamente por la empresa suministradora". A veces el procedimiento de corrección es manual, pero sin plantilla. En 2.11.2.1 se pregunta por el tamaño de las muestras. No aparece entre las opciones "varios estudios con muestras pequeñas". En baremos se puntúa la calidad de las normas y el tamaño de los grupos, pero ¿cómo se tiene en cuenta la aplicación de estrategias, como "continuous norming" (Zachary & Gorsuch, 1985), que palian el problema del reducido tamaño de los grupos normativos?

En relación al proceso de revisión

Muñiz et al. (2011), al hacer balance de la primera evaluación, reconocen que el proceso a seguir para hacer la revisión plantea algunas dudas, y seguramente no es independiente de lo anterior el hecho de que los distintos países sigan procedimientos distintos. En nuestro país, como se ha descrito anteriormente, el procedimiento de revisión se parece mucho a cómo se revisan los artículos en las revistas científicas, pero hay algunas diferencias.

La revisión de un test no comercial plantea dificultades propias, no siendo la menor de ellas la determinación de sobre qué artículos/informes han de hacer los revisores la evaluación, al carecer de manual. El primer test no comercial revisado ha sido el EPV-R y lo ha sido en esta segunda revisión. El coordinador optó por pedir ayuda a los autores, que por cierto la prestaron muy amable y eficazmente. Queda la duda de cómo proceder si alguna vez no fuera así. Tiene sentido que la CT ponga en marcha la revisión del test que considere oportuno y la haga pública, pero convendría establecer un protocolo de actuación en estos casos. Podría indicar, por ejemplo, si se ha de informar y pedir colaboración al autor del test y quién ha de hacerlo, cómo determinar sobre qué documentos se ha de basar la revisión, si hay que explicitar las razones por las que el test ha sido elegido, entre otros posibles contenidos.

La ciencia, a veces, no casa del todo bien con el negocio. Si el manual informa de muchos detalles psicométricos, el precio del test es más alto, por el esfuerzo de hacerlo y por lo que cuesta el mismo manual (más páginas). Pudiera entonces ocurrir que seguir las recomendaciones de mejora que los revisores proponen redunde en dificultades para su comercialización. Una posible solución quizás sea que los editores ofrezcan la información fundamental en el manual y la más sofisticada en la web. Otro ejemplo de esto mismo se plantea cuando el editor no publica en el manual información psicométrica relevante (por ejemplo, los pesos de los ítems en los factores). Los revisores puntúan negativamente que no se ofrezcan estos datos, a pesar de que la informa-

ción existe. Una posible solución a esto último es que los editores suministren a los revisores, junto con el test, la información disponible requerida por el CET que no aparezca en el manual. Esto habría de hacerse salvaguardando la confidencialidad de dicha información y el anonimato del proceso de revisión. Un tercer ejemplo tiene que ver con las consecuencias de la revisión: de los tests revisados se explicitan sus puntos fuertes, pero también sus puntos débiles. Estas informaciones no están disponibles en el caso de los tests no revisados, quedando la duda de si éstos últimos pudieran resultar mejor parados. Evers et al. (2010) afirman que en Holanda se ha establecido la idea de que una buena práctica en el uso de los tests es aplicar los que hayan recibido buenas evaluaciones en el proceso de revisión. No podemos estar más de acuerdo y lo esperable es que algo similar suceda en nuestro país cuando el proceso de revisión vaya progresando y sean más y más los tests revisados.

Relacionado con lo indicado al comienzo del párrafo anterior cabe hacerse la siguiente pregunta: ¿Tiene sentido recomendar análisis psicométricos complejos (pruebas de invarianza, estudios de funcionamiento diferencial, indicadores de precisión de cada medida...) y solicitar que se aporten detalles (por ejemplo, información del ajuste a los modelos) que probablemente no entienden la mayoría de los psicólogos aplicados? Los usuarios lo que seguramente reclaman son ayudas para la interpretación de las puntaciones. ¿Debiera el proceso de revisión incorporar más activamente el punto de vista de los profesionales? Los revisores somos en su casi totalidad académicos, ¿no tendremos el sesgo de evaluar un test como evaluamos los artículos, cuando el objetivo y público objetivo son distintos? Como indica Elosua (2012), la Psicometría moderna está progresando muy deprisa y la distancia entre los desarrollos actuales y los que se vienen aplicando en las tesis, artículos, manuales de tests, etc., por los no expertos en Psicometría suele ser considerable. Recientemente, en esta misma revista, Ponsoda (2010) coordinó el monográfico "Metodología al servicio del psicólogo", cuyo propósito era acercar al profesional algunos de los desarrollos psicométricos modernos, que muy probablemente no habían estudiado anteriormente, como el sesgo de los ítems y tests, el análisis factorial confirmatorio, los modelos de ecuaciones estructurales, las conceptos recientes de fiabilidad y validez, las nuevas teorías de los tests, los formatos innovadores de tests e ítems... A nuestro modo de ver, los nuevos desarrollos psicométricos que los revisores recomiendan ayudan a mejorar los tests, pues de su aplicación resultarán nuevas evidencias de validez, indicadores alternativos de la fiabilidad de la prueba, indicadores de la precisión de las medidas individuales, de obvio interés cuando estamos interesados en la evaluación individual frente a la colectiva, etc. Lo anterior no es contradictorio con que el manual satisfaga además las necesidades y exigencias del usuario y facilite la aplicación correcta y cómoda del test. En este mismo sentido, sería interesante incorporar la visión del usuario en el proceso de revisión, pero obviamente no podría ser pidiéndole responder al CET. Habría que pensar en algún procedimiento que informe de su satisfacción con el test, que permita conocer sus puntos fuertes y débiles desde la perspectiva del usuario y no del experto. Lo anterior parece difícilmente integrable en el proceso de revisión actual. Información de este tipo se obtiene con las encuestas sobre la opinión sobre los tests (Muñiz y Fernández-Hermida, 2000, 2010), si bien obviamente no es específica de un test particular. En cuanto a la tercera pregunta, sí creemos que hay un cierto riesgo de hacer las revisiones como hacemos las de los artículos, dada la escasa experiencia que tenemos todos en revisar tests. Es tarea del coordinador, en sus interacciones con los revisores, hacerles ver que, en efecto, el objetivo de la revisión no es el mismo que el de los artículos científicos. Por tanto, sus recomendaciones deben ceñirse a los asuntos que mejoren la prueba, aporten nuevas evidencias de validez... y no a propuestas que pudieran eventualmente mejorar el conocimiento de los procedimientos psicométricos aplicados o del constructo que el test mide.

Puede hacerse también algún comentario en relación al papel del coordinador. En las dos revisiones anteriores cada revisor ha recibido el test, regalado por el editor. Lo cierto es que el coordinador también necesita las pruebas sobre las que ha de hacer los informes, para añadir a las dos revisiones una tercera si lo considera adecuado, para aclarar las discrepancias entre revisores y para hacer alguna eventual comprobación sobre lo que autores y editores proponen cambiar en la revisión que se les envía, antes de la evaluación final. Una posible solución es que los editores y/o la CT proporcionen al centro de trabajo del coordinador las pruebas a revisar que el centro no tenga y no pueda adquirir.

CONCLUSIONES

En primer lugar, conviene destacar que el proceso de evaluación de tests iniciado en 2010 continúa y va consolidándose; no obstante, para que esta consolidación resulte más útil, sería conveniente que se produjera un rápido incremento del número de tests revisados. Hasta el momento se han revisado 22 tests. Uno de ellos es un test no comercializado, elegido por su repercusión social. El test EPV-R, de Echeburúa et al. (2010), se ha incorporado al protocolo que siguen en muchas comisarías de policía, tras una denuncia de agresión contra la mujer, para la predicción de riesgo de violencia grave contra la pareja. Anteriormente, los policías decidían subjetivamente las medidas de protección que debían tomar; con la aplicación del test, las medidas de protección son las establecidas para el nivel de riesgo que se sigue de su aplicación.

El nivel medio de los tests revisados es, en términos absolutos, bueno (4, en una escala que va de 1 a 5); y casi coincide con el obtenido en la primera evaluación. Sucesivas evaluaciones mostrarán si es ésta la calidad media de las pruebas editadas en España, o si, a medida que más y más tests son evaluados, la media cambia. En las dos evaluaciones hemos comprobado que sólo un test da información detallada del sesgo o funcionamiento diferencial de los ítems. Están por debajo de la media en ambas evaluaciones 3 características: "Análisis de los ítems", "Validez predictiva", y "Fiabilidad entendida como estabilidad". En el apartado de puntos a mejorar, se sugiere en varias ocasiones que se informe en el manual de las propiedades individuales de los ítems y los criterios de selección para configurar el test final, que se proporcione más evidencia de la capacidad de la prueba para predecir criterios relevantes en sus campos de aplicación, y que se procure incorporar en nuevas ediciones del test estudios de fiabilidad test-retest, que informen de la estabilidad de las puntuaciones.

Los revisores han advertido algunas otras carencias, como las siguientes: la escasa justificación de los puntos de corte aportados para la interpretación de las puntuaciones, la no obtención de indicadores de la precisión de la puntuación de cada evaluado sino del test en su conjunto, la escasez de estudios de funcionamiento diferencial de ítems y tests, y la escasez de estudios que aporten evidencias y justifiquen ciertos usos esperables de las puntuaciones. La botella se puede ver también medio llena. En general se ha cuidado mucho la correcta inserción del constructo que el test mide en la teoría psicológica. Se aprecia el uso en varios tests de la teoría de la respuesta al ítem, que proporciona por cierto indicadores de la precisión de cada medida. Se ha realizado algún estudio de funcionamiento diferencial; se han aplicado en algunos tests análisis factoriales confirmatorios y modelos de ecuaciones estructurales; se han utilizado desarrollos recientes para la construcción de baremos que permiten obtener muchos distintos, evitando que el tamaño requerido de la muestra total sea demasiado grande; en algunos tests se han contemplado "acomodaciones", o cambios a hacer en la administración de la prueba cuando se aplica a individuos con alguna característica especial, que posibiliten una adecuada interpretación de las puntuaciones; varios tests proporcionan sistemas automáticos de corrección e interpretación de puntuaciones; y, por último, en algunas de las pruebas revisadas se cuida la representación de las muestras normativas, incluyendo muestras no incidentales, y se aplican procedimientos de validación cruzada para evitar indicadores psicométricos artificialmente altos. En fin, es una buena noticia que la distancia entre la teoría y la práctica, aludida en el apartado anterior, va reduciéndose.

La evaluación de la calidad de los tests necesita del CET por varias razones. Facilita la tarea de los revisores y de los autores y editores del test, al indicar exactamente qué se va a valorar y cómo. Permite a la vez una evaluación cuantitativa y cualitativa de las características relevantes a tener en cuenta cuando se quiere determinar la calidad de un test. Facilita la comparación de resultados entre tests distintos y tandas de evaluación. Evidentemente, lo anterior no excluye que requiera revisiones. En general, el CET se podría mejorar añadiendo un glosario con los términos susceptibles de diferentes interpretaciones, y aclaraciones y ejemplos sobre las preguntas problemáticas. Alternativamente, podría también contemplarse la inclusión de algún procedimiento de búsqueda de consenso entre revisores.

El CET está inspirado en el modelo de evaluación de la Federación Europea de Asociaciones de Psicólogos Profesionales y este modelo está siendo modificado en profundidad (Evers et al., 2010). La incorporación de algunas de estas modificaciones y la respuesta a las dificultades de aplicación comentadas en este artículo y en Muñiz et al. (2011) sugieren que convendría plantearse su modificación. Todo indica que en los próximos meses se publicará una nueva edición de los "standards" (AERA, APA y NCME, 1999). Compartiendo con Elosua

(2012) la idea de que hemos de tenerlos muy en cuenta, la próxima aparición de los nuevos es otra razón para modificar el CET.

En cuanto al proceso de revisión en su conjunto, nuestra impresión es que funciona razonablemente bien; lo que no excluye que pueda introducirse algún cambio que se considere oportuno. Se entrega a los revisores una cantidad simbólica de 50 euros. Algunos revisores han preferido no cobrarla, pues consideran que, como ocurre con la revisión de artículos, son tareas que no deben ser remuneradas. Siguiendo el modelo de las revistas científicas que nombran al editor y comité editorial por un periodo de 2 o 3 años, se podría considerar si es apropiado o no que el coordinador y los revisores evalúen más de una tanda de tests, durante dos o tres años, e informen de los resultados de la revisión al final del periodo.

AGRADECIMIENTOS

Queremos agradecer su colaboración a los miembros de la Comisión de Tests: José Ramón Fernández Hermida, Ana Hernández Baeza, Miguel Martínez García, Milagros Antón López, Viviana Gutman Mariach, y, en especial a su Presidente, José Muñiz Fernández, por su continua ayuda. Queremos reiterar también nuestro agradecimiento a los revisores por su entusiasta y excelente colaboración.

REFERENCIAS

American Educational Research Association, American Psychological Association, y National Council of Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Echeburúa, E., Amor, P.J., Loinaz, I. y Corral, P. (2010). Escala de Predicción de Riesgo de Violencia Grave contra la pareja – Revisada (EPV-R). *Psicothema, 22(4),* 1054-1060.

Elosua, P. (2012). Tests publicados en España: Usos, costumbres y asignaturas pendientes. *Papeles del Psicólogo, 33(1),* 12-21.

Evers, A., Sijtsma, K., Lucassen, W. & Meijer, R.R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure and results. *International Journal of Testing, 10,* 295-317.

Muñiz, J. y Fernández-Hermida, J.R. (2000). La utilización de los tests en España. *Papeles del Psicólogo, 76,* 41-49.

Muñiz, J. y Fernández-Hermida, J.R. (2010). La opinión de los psicólogos españoles sobre el uso de los tests. *Papeles del Psicólogo, 31(1),* 108-121.

Muñiz, J., Fernández-Hermida, J.R., Fonseca-Pedrero, E., Campillo-Álvarez, A. y Peña-Suárez, E. (2011). Evaluación de tests editados en España. *Papeles del Psicólogo, 32(2),* 113-128.

Ponsoda, V. (2010). Metodología al servicio del psicólogo. *Papeles del Psicólogo, 31(1),* 2-6.

Prieto, G. y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 77,* 65-71.

Zachary, R. A. & Gorsuch, R. L. (1985). Continuous norming: Implications for the WAIS-R. *Journal of Clinical Psychology, 41,* 86–94.