Papeles del Psicólogo / Psychologist Papers, 2021 Vol. 42(1), pp. 1-9
https://doi.org/10.23923/pap.psicol2020.2937
http://www.papelesdelpsicologo.es
http://www.psychologistpapers.com

OCTAVA EVALUACIÓN DE TEST EDITADOS EN ESPAÑA: UNA EXPERIENCIA PARTICIPATIVA

EIGHTH REVIEW OF TESTS PUBLISHED IN SPAIN: A PARTICIPATIVE EXPERIENCE

Carme Viladrich[1], Eduardo Doval[1], Eva Penelo[1], Joan Aliaga[1], Albert Espelt[1,2], Rebeca García-Rueda[1] y Ariadna Angulo-Brunet[1]
[1]*Universitat Autònoma de Barcelona.* [2]*CIBER de Epidemiología y Salud Pública*

La Comisión de Test del Consejo General de la Psicología en España promueve anualmente la revisión de la calidad de diferentes test publicados. Este trabajo tiene un doble objetivo: a) presentar los resultados de la octava edición y b) considerar la aportación de la universidad en dicho proceso. En esta edición participaron 10 especialistas, 332 estudiantes y siete profesores, adaptándose el protocolo estándar de revisión al formato aprendizaje-servicio. En cuanto a los resultados, la calidad de los 11 test evaluados fue adecuada (promedio de 3,9 puntos en una escala 1-5) y similar a años anteriores (r = 0,90). El desarrollo y la baremación fueron puntos fuertes, mientras que se proponen mejoras en otros aspectos. El aprendizaje-servicio contribuyó a la diversificación de voces en el proceso observándose una calidad similar entre los informes del estudiantado y los emitidos por especialistas y un grado de acuerdo esperable (r = 0,67) entre ellos. Concluimos que el presente proyecto ha permitido identificar la oportunidad de profundizar en el uso de lenguaje compartido para fortalecer la comunicación entre las casas editoriales, la comisión promotora del modelo español de revisión de test, y las personas usuarias de los test, particularmente si se trata de principiantes.
***Palabras clave**: Evaluación de test, Calidad de los test, Psicometría, Aprendizaje-servicio.*

Every year, the Test Commission of the Spanish Psychological Association promotes the assessment of the test quality of several published tests. The aim of the present study is two-fold: a) to present results for the eighth review, and b) to consider the contribution of the universities in this process. Ten experts, 332 students, and seven professors participated in this edition and the standard protocol for review was aligned towards a service-learning format. For the 11 tests assessed, results showed an adequate quality (average of 3.9 points on a 1-5 rating scale) similar to previous years (r = .90). The strengths were test development and standardization, and a number of proposals for improving other sections were suggested. The service-learning approach contributed to the diversification of voices in the process with students' and experts' reports showing similar quality and an expected level of agreement (r = .67). We conclude that this project has helped to identify the opportunity to further deepen the use of shared language in order to strengthen the communication between the test publishers, the promoters of the Spanish model of test assessment, and the test users, especially in the case of beginners.
***Key words:** Test review, Test quality, Psychometrics, Service-learning.*

La revisión de test que promueve la Comisión de Test del Consejo General de la Psicología de España (COP, http://cop.es) en colaboración con la Comisión de Test de la *European Federation of Psychologists' Associations* (EFPA, http://efpa.eu; Evers et al., 2013) se inició en 2011 (Muñiz et al., 2011) y desde entonces se han publicado 84 revisiones al ritmo aproximado de una edición anual. Con esta iniciativa se pretende dar respuesta a las necesidades de información independiente por parte de la profesión (e.g., Hidalgo y Hernández, 2019), así como de formación (e.g., Fonseca-Pedrero y Muñiz, 2017), que son más relevantes cuando se utilizan los test para tomar decisiones con consecuencias importantes para las personas evaluadas (e.g., Hernández et al., 2015).

Entre los actores que intervienen en el proceso, la Comisión de Test del COP, formada por profesionales del ámbito académico y de las casas editoriales, ha asumido la función de priorizar los test a evaluar en cada edición. A su vez, el perfil mayoritario de participantes en las revisiones ha sido el de personas que ocupan posiciones senior en distintas especialidades académicas en los ámbitos de la psicología y la educación. Una vez consolidado el proceso, se ha hecho explícita la voluntad de incluir nuevas voces en la revisión (Elosua y Geisinger, 2016), particularmente las que hablan desde el ámbito profesional y desde posiciones junior (Fonseca-Pedrero y Muñiz, 2017). Un paso más en esta dirección sería la ampliación a estudiantes que se incorporarán en breve a la profesión, bajo la tutoría de su profesorado. Además de integrar una nueva voz en el proceso, esta estrategia daría la oportunidad al profesorado de compaginar las tareas de revisión y de formación, al menos en algunas especialidades académicas.

Recibido: 27 abril 2020 - Aceptado: 8 junio 2020
Correspondencia: Eva Penelo. *Facultat de Psicologia. Universitat Autònoma de Barcelona. C. de la Fortuna s/n. 08193 Bellaterra (Cerdanyola del Vallès). España.*
E-mail: eva.penelo@uab.cat

El potencial del modelo europeo de revisión de test como herramienta de formación ha sido ampliamente reconocido por parte del profesorado universitario (Hidalgo y Hernández, 2019; Vermeulen, 2019). Así lo vimos también en la asignatura de Psicometría de la *Universitat Autònoma de Barcelona* (UAB, http://uab.cat) cuando a partir del curso 2011/12 implementamos un proyecto de aprendizaje basado en problemas utilizando el modelo de evaluación de test promovido por el COP. El estudiantado redacta un informe de evaluación de un test psicológico, educativo o logopédico completando el Cuestionario de Evaluación de Test Revisado (CET-R; Hernández, et al., 2016) y lo defiende oralmente como parte de las evidencias de aprendizaje que presenta (Doval et al., 2013; Viladrich, Doval, Aliaga et al., 2014). Durante la última década, hemos evaluado un total de 91 test entre los disponibles en nuestra docimoteca, hemos presentado datos favorables de la validez de sus revisiones en comparación con las de especialistas (Viladrich, Doval y Penelo, 2014), hemos estudiado el efecto de la adhesión temprana al proyecto en los resultados académicos (Espelt et al., 2016), y hemos contribuido a la revisión del modelo español (Hernández et al., 2016), además de colaborar individualmente como revisores en distintas ediciones. Durante el curso académico 2019/20 hemos aceptado el reto de liderar la octava edición de la evaluación de test editados en España. Para ello, hemos adaptado la metodología docente que veníamos utilizando a un formato de aprendizaje-servicio (ApS, Redondo-Corcobado y Fuentes, 2018) que ha salido del ámbito de la universidad para dirigirse al conjunto de la comunidad profesional (Viladrich et al., 2019, 2021).

En esta edición, la Comisión de Test del COP nos encargó que revisáramos 11 test encaminados a la medida de la inteligencia, las aptitudes verbales, y la personalidad, publicados entre los años 2006 y 2019. Concretamente, se trata de los seis niveles del test BADyG, y de los test BRIEF-P, CELF-5, MCMI-IV, PECO y TONI-4. Pueden verse más detalles de todos ellos en la Tabla 1. En consecuencia, el primer objetivo de este artículo es exponer y discutir los resultados de calidad de los 11 test sometidos a evaluación en la octava edición de la evaluación de test editados en España. El segundo objetivo es exponer y discutir la aportación de la universidad en relación con dos roles novedosos: como proponente de los test a evaluar y como participante en el proceso de revisión a través del estudiantado de psicología bajo la tutoría de su profesorado.

MÉTODO

Participantes

Participaron 332 estudiantes (78,9% de sexo femenino que formaron 69 equipos de trabajo y el equipo de profesorado de la asignatura de Psicometría, que es obligatoria de tercer curso del grado en Psicología de la UAB. Por otra parte, participaron seis revisoras y cuatro revisores especialistas en psicometría, salud o educación provenientes de distintas instituciones españolas (véase parte superior izquierda de la Tabla 2) y un número no concretado de personas de cada casa editorial.

Instrumentos

CET-R. Los criterios de calidad del modelo español de evaluación de test se reflejan en el CET-R (Hernández et al., 2016), que está formado por tres apartados. En el primero, se describen las características del test; en el segundo, se evalúan sus propiedades; y en el tercero, se resumen todas las valoraciones. Las propiedades de los test se valoran contestando preguntas cerradas con cinco categorías de respuesta ordenadas desde insuficiente a excelente (10 preguntas sobre el desarrollo del test, 18 sobre validez, 14 sobre fiabilidad y nueve sobre

TABLA 1
TESTS EVALUADOS EN LA OCTAVA EDICIÓN

Acrónimo	Nombre	Editorial	Año de publicación
BADyG/i	Batería de Actividades mentales Diferenciales y Generales, Nivel infantil	CEPE, S.L.	2019
BADyG/E1-r	Batería de Actividades mentales Diferenciales y Generales, Nivel E1 renovado	CEPE, S.L.	2019
BADyG/E2-r	Batería de Actividades mentales Diferenciales y Generales, Nivel E2 renovado	CEPE, S.L.	2019
BADyG/E3-r	Batería de Actividades mentales Diferenciales y Generales, Nivel E3 renovado	CEPE, S.L.	2019
BADyG/M-r	Batería de Actividades mentales Diferenciales y Generales, Nivel M renovado	CEPE, S.L.	2019
BADyG/S-r	Batería de Actividades mentales Diferenciales y Generales, Nivel S renovado	CEPE, S.L.	2019
BRIEF-P	Evaluación Conductual de la Función Ejecutiva- Versión infantil	TEA Ediciones	2016
CELF-5	Evaluación Clínica de los Fundamentos del Lenguaje, 5	Pearson Educación	2018
MCMI-IV	Inventario Clínico Multiaxial de Millon, IV	Pearson Educación	2018
PECO	Prueba para la Evaluación del Conocimiento Fonológico	EOS	2006
TONI-4	Test de Inteligencia No Verbal, 4	TEA Ediciones	2019

interpretación de las puntuaciones). Además de estar basadas en una rúbrica, estas valoraciones se argumentan en varias preguntas abiertas. Puesto que no todos los test requieren las mismas evidencias de calidad, el modelo se flexibiliza sometiendo al juicio de quién evalúa un test concreto la aplicabilidad de cada tipo de evidencia, particularmente en los apartados de fiabilidad e interpretación de puntuaciones.

Rendimiento y satisfacción del estudiantado. Se proporcionan de forma estándar para todas las asignaturas de la universidad. Se valora el porcentaje de retención y el de éxito (UAB, n.d.-a), así como percepción de puntos fuertes y puntos débiles de la asignatura (UAB, n.d.-b).

Procedimiento

En la Figura 1 se representa el procedimiento que consiste en obtener dos revisiones independientes, conciliadas por la editora que, a continuación, considera los comentarios de las casas editoriales antes de redactar y difundir el informe definitivo y el proceso de evaluación. Véase Gómez (2019) para más detalles sobre protocolo estándar. Por nuestra parte, hemos desarrollado un protocolo específico para aplicarlo al proyecto académico de ApS. Cada una de las siete personas que firman este trabajo actuó como editor/a de uno o dos de los test a evaluar hasta cubrir los 11 que formaron el encargo. La primera autora, además, canalizó todos los flujos de información con la Comisión de Test, con

TABLA 2
PARTICIPANTES EN LA OCTAVA EVALUACIÓN DE TESTS

Revisor/a profesional (Filiación)	**Editor/a-Tutor/a (Universitat Autònoma de Barcelona)**
Alejandro Veas Iniesta (Universidad de Alicante)	Albert Espelt
Ana Isabel González Contreras (Universidad de Extremadura)	Ariadna Angulo-Brunet
Gerardo Aguado Alonso (Universidad de Navarra)	Carme Viladrich
Maria Dolores Gil Lario (Universitat de València)	Eduardo Doval
Maria Dolores Prieto (Universidad de Murcia)	Eva Penelo
Marijo Garmendia Txapartegi (Zarauzko Berritzegunea)	Joan Aliaga
Miguel Angel Carbonero (Universidad de Valladolid)	Rebeca García-Rueda
Montse Bartroli Checa (Agència de Salut Pública de Barcelona)	
Natalia Hidalgo-Ruzzante (Universidad de Granada)	
Rafael Martínez Cervantes (Universidad de Sevilla)	
Revisor/a estudiante (Universitat Autònoma de Barcelona)	
Ainoa Barreiro Escobar	Laura Saavedra García
Alba Rodríguez-Delgado	Layla Ishak-Tello
Aleix Jané-Alsina	Mar Viniegra Pintado
Anna Orts	Maria Peiro
Cristina Fuste-i-Valentí	Maria Silva Pereira
Carmen María Segura Sanchez	Marian Granados-Gamito
Daniel Steinherr Zazo	Marina Clivillé Domingo
Dunia Hanafi-Alcolea	Marta Valera-Guiot
Fátima Zarfani Azarfane	Meritxell Bagué Solé
Fernando Mengual-Rodas	Meritxell Barroso Cantero
Gemma Casimiro Fernandes	Mireia Gamez-Broto
Gemma Lapeira-Casé	Natalia Llobet-Vallribera
Isaac Pardo-Niño	Núria Coma Bravo
Joan Martínez-Vidal	Oriol Martín-Corella
Judit Reyes Griñena	Paula Jimenez-Ventura
Judith Moya	Queralt Mas-Jarque
Júlia Bartra Pallarès	Raquel Villar Mateo
Júlia Carrasco Hernández	Roser Bigorra Fargas
Ksenia Ouzioumova	Sílvia Solano Selvas
Laia Cervigón Moreno	Xenia Pla-Ruiz

las revisoras, los revisores y con las casas editoriales y supervisó todos los informes. Durante el curso académico, en cada grupo de prácticas de Psicometría se evaluó uno de los test en forma de competición entre cuatro o cinco equipos formados por entre tres y seis personas. Cada equipo desarrolló un borrador, recibió los comentarios de su tutor o tutora, y redactó después el informe final. El informe ganador de cada test fue considerado como la Revisión 2 (véanse las autorías en la parte inferior de la Tabla 2). En caso que el premio se considerara desierto, el profesorado tutor asumía este papel. Posteriormente, el profesorado llevó a cabo dos sesiones de discusión para fijar una línea editorial común y valorar el proyecto.

En cuanto a las precauciones éticas, la adaptación al formato ApS fue como sigue. Cada miembro del equipo de profesorado aceptó participar en el proyecto antes de empezar el curso; como alternativa, podía restringir su actividad a la usual de tutorizar a sus estudiantes. Para el estudiantado fue una actividad obligatoria y evaluable como lo venía siendo en la última década. Se les informó del proyecto durante la primera sesión de clase y de forma permanente en el aula virtual. Cada estudiante reconoció mediante su firma haber leído el documento sobre las leyes de copyright que afectan a los materiales. Por otra parte, cada firmante de un proyecto ganador dio por escrito su consentimiento para hacer uso de su texto como Revisión 2 y para publicar su nombre en los procesos de difusión. La gratificación económica que ofrece el COP a la editora y al Revisor 2 se ingresó en un fondo de la UAB destinado a sufragar gastos de la asignatura de Psicometría.

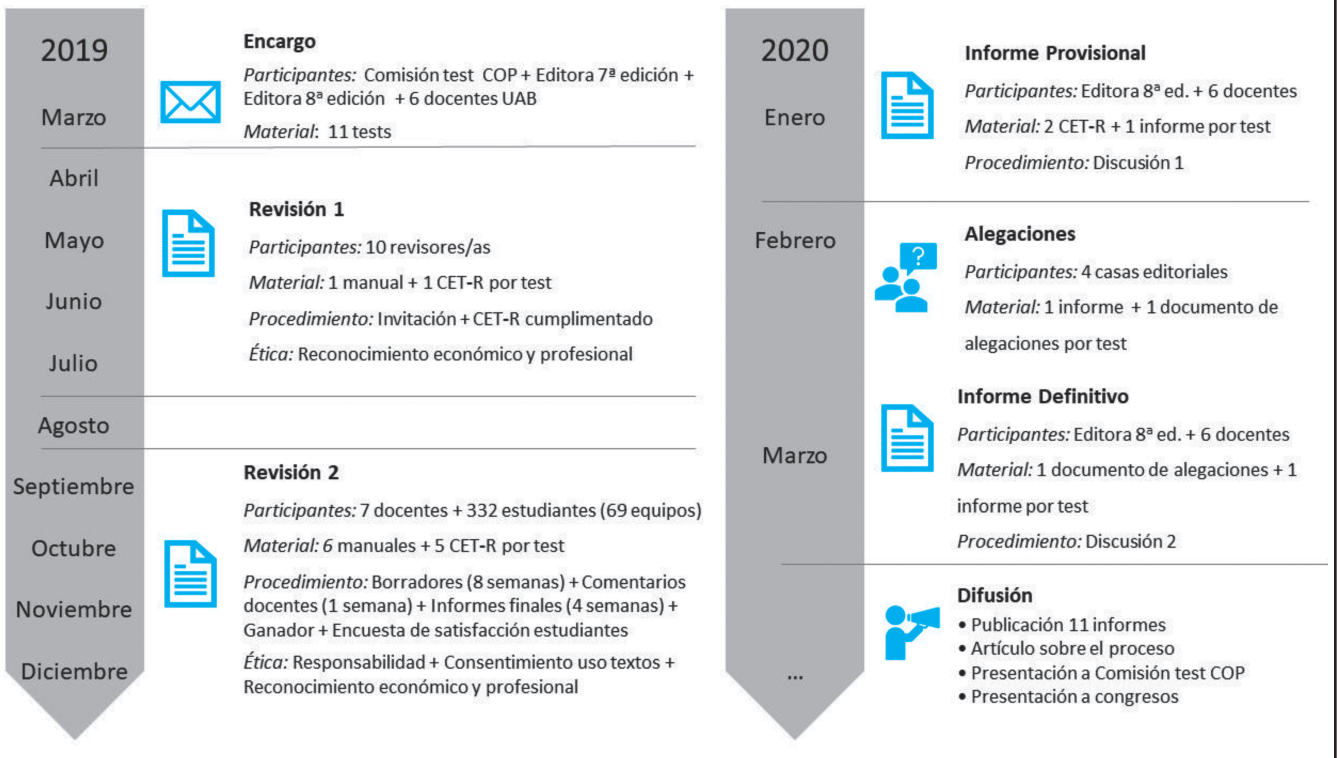
RESULTADOS Y DISCUSIÓN

Calidad de los Test Sometidos a Evaluación

Los informes de revisión de los 11 test sometidos a evaluación son el resultado principal de este trabajo y pueden consultarse y descargarse en el sitio web del COP, dentro del apartado correspondiente al año 2019 (https://www.cop.es/index.php?page=evaluacion-tests-editados-en-espana). A modo de resumen, en la Tabla 3 pueden observarse las puntuaciones de cada uno de los 11 test evaluados en los aspectos relativos al desarrollo del test, las evidencias de validez, la estimación de la fiabilidad de las puntuaciones, y la baremación e interpretación de puntuaciones.

En primer lugar, destacamos que los test valorados en esta edición están desarrollados y baremados de forma muy correcta, tal y como se desprende de la primera y última categorías en la Tabla 3. En todos los manuales se presenta una fundamentación teórica entre buena y excelente y lo mismo sucede con el proceso de adaptación. Aunque con

FIGURA 1
PROCEDIMIENTO DE REVISIÓN DE LOS TESTS EN LA OCTAVA EDICIÓN



algo más de variabilidad, también los materiales y la documentación de los test se han valorado en este mismo rango, así como la gran mayoría de los estudios de baremación. Estos resultados se mantienen a una altura similar o superior a los obtenidos en las otras revisiones basadas en el CET-R (Gómez, 2019; Fonseca-Pedrero y Muñiz, 2017; Hidalgo y Hernández, 2019) y, en conjunto, ilustran la robustez de los test que las casas editoriales someten a evaluación año tras año en estos aspectos. Cabe destacar que la presentación de datos de análisis de ítems no siempre alcanza valores aceptables según la rúbrica del CET-R. En este sentido, recuperamos aquí la sugerencia de Ponsoda y Hontangas (2013) de que en los manuales se podrían mencionar materiales suplementarios que estuvieran disponibles en la página web de la editorial, como ya se hace, por ejemplo, con la batería de test BADyG. Respecto a la interpretación de puntuaciones, la mejora que juzgamos más importante sería que en todos los manuales se justificaran los puntos de referencia que se publican a modo de ayuda para la toma de decisiones. Así reza la recomendación del CET-R (apartado 2.13.2.1), y un ejemplo de buena práctica son los datos de sensibilidad y especificidad que se publican en el manual del test MCMI-IV. Como mínimo, en todos los manuales habría que hacer un esfuerzo por aclarar que el hecho de que una persona ocupe una posición atípica en relación con su grupo normativo no tiene en sí mismo significación clínica. Se contribuiría así a corregir posibles malos usos que pueden ser razonablemente anticipados (*American Educational Research Association* [AERA] et al., 2014, estándar 7.1).

La información contenida en el apartado de validez ha sido calificada mayoritariamente como adecuada, lo que significa que el uso de los test evaluados ha merecido confianza a quienes han revisado, aunque sin duda es el apartado con mayor margen de mejora. Una primera oportunidad de mejora sería asociar a cada uno de los usos propuestos del test aquellas evidencias de validez que lo sustentan. Así se recomendaba en la revisión firmada por Elosua y Geisinger (2016), así está contemplado en la introducción al apartado 2.11 del CET-R, y en ello se insistía con la propuesta de Gómez (2019) de reconsiderar la valoración de la etiqueta "No se aporta información" en el CET-R. De hecho, escribir en el manual de un test la afirmación "el test TAL puede utilizarse para evaluar la característica X" requiere un soporte distinto que escribir la afirmación "el test TAL puede utilizarse para detectar dificultades de una persona en la característica X, para diseñar un plan de intervención encaminado a su mejora, y para hacer el seguimiento de su evolución en los ámbitos clínico, educativo, social y legal". La diferencia radica en que la primera afirmación deja bajo la responsabilidad de la persona usuaria del test el uso concreto que hará de esa evaluación y, por tanto, la responsabilidad de sustentar dicho uso (AERA et al., 2014, estándar 9.4). En cambio, con la segunda afirmación se promueve explícitamente el test para diversos usos específicos, por lo que la responsabilidad de apoyar cada uso concreto recae en la casa editorial (AERA et al., 2014, estándares 5.0 y 7.1).

Otra mejora se conseguiría si, antes de presentar resultados de validez, se especificaran con toda claridad las hipótesis concretas que se van a comprobar y cuáles son los resulta-

TABLA 3
RESUMEN DE LAS CALIFICACIONES DE LOS TEST EVALUADOS EN LA OCTAVA EDICIÓN

Característica	BADyG						BRIEF	CELF	MCMI	PECO	TONI	Media 2019
	i	E1-r	E2-r	E3-r	M-r	S-r	P	5	IV	4		(anteriores*)
Desarrollo: Materiales y documentación	4,5	4,5	4,5	4,5	4,5	4,5	4,8	4	4,3	3	4,5	4,3(4,3)
Desarrollo: Fundamentación teórica	5	5	5	5	5	5	4,5	4	4	4	5	4,7(4,1)
Desarrollo: Adaptación	-	-	-	-	-	-	5	5	4,5	-	-	4,8(4,3)
Desarrollo: Análisis de ítems	4	4	4	4	4	4	4	2	2	4	4	3,6(3,8)
Validez: contenido	4	4,5	5	5	5	4,5	3,3	3,5	3	3	3,5	4,0(3,8)
Validez: relación con otras variables	3,5	3,4	2,7	3,7	3,7	3	3,3	3,4	3,4	2,6	3,9	3,3(3,6)
Validez: estructura interna	2,5	3	3	3	3	3	2,5	-	-	2,5	5	3,1(3,7)
Validez: análisis del DIF	-	-	-	-	-	-	-	-	-	-	5	-
Fiabilidad: equivalencia	-	-	-	-	-	-	-	-	-	-	3	-
Fiabilidad: consistencia interna	4	4	5	5	5	4	5	4,5	4	3,5	3,5	4,3(4,2)
Fiabilidad: estabilidad	-	-	-	-	-	-	4	2,5	3	-	3	3,1(3,5)
Fiabilidad: TRI	-	-	-	-	-	-	-	-	-	-	-	-
Fiabilidad inter-jueces	-	-	-	-	-	-	-	-	-	-	-	-
Baremos e interpretación de puntuaciones	4,7	4,3	4,7	4,3	4,7	4,7	3,7	3,3	4	2,3	4,3	4,1(4,1)

Nota. 1: Inadecuado, 2: Adecuado con carencias, 3: Adecuado, 4: Bueno, 5: Excelente, -: No pertinente o No se aportaron datos. *: Puntuación media entre las medias de las ediciones desde 2010 a 2018.

dos que se considerarán evidencia favorable teniendo en cuenta la fundamentación teórica, los resultados obtenidos en investigaciones anteriores y los usos previstos del test (e.g., Ziegler, 2014). En este sentido, la cadena hipótesis-resultados-conclusiones debería quedar clara en relación con todas las cargas factoriales, todos los coeficientes de correlación y todos los tamaños del efecto que se publican, incluso si se presentan dentro de tablas o como resultados previamente publicados. Ello no impide incorporar muchas variables en unas pocas hipótesis, como es el caso de las que se hacen en el análisis factorial o en el diseño de matrices multirasgo-multimétodo; al contrario, este sería precisamente un formato muy recomendable.

La tercera oportunidad de mejora en el ámbito de la validez consistiría en incorporar más información sobre la equivalencia y la equidad en el uso de los test. De forma consistente, en las revisiones anteriores se ha animado a hacerlo incrementando la publicación de estudios sobre funcionamiento diferencial de los ítems (DIF, por sus siglas en inglés; e.g., Fonseca-Pedrero y Muñiz, 2017; Gómez, 2019; Hidalgo y Hernández, 2019). Con todo, en la presente edición su uso se mantiene a un nivel parecido o menor, ya que se ha aportado un único análisis DIF, el relativo al test TONI-4. Por nuestra parte, valoramos el coste que tiene hacer este tipo de pruebas, incluyendo el riesgo de sobreestimar la presencia de DIF. Por ello, creemos que ha llegado el momento de recomendar un paso más asequible y, sin embargo, fundamental, para dar a la equidad la importancia que se le reconoce actualmente tanto en la sociedad como en los textos normativos (AERA et al., 2014; COP, 2015a, 2015b; Asociación Española de Normalización y Certificación [AENOR], 2013). Concretamente, sugerimos incluir en los manuales un apartado específico dedicado a aportar información sobre la flexibilidad del test para afrontar la diversidad funcional, lingüística, neurológica o social de las personas potencialmente evaluables. Esta información puede recabarse mediante análisis DIF, pero únicamente cuando los grupos son grandes y las hipótesis bien definidas. En cambio, es mucho más factible obtener información relevante mediante consultas a especialistas y a personas integrantes de los grupos minorizados (AERA et al., 2014, estándar 3.11). Esta información puede y debe obtenerse y compartirse de forma rigurosa (Levitt et al., 2018). Un rigor que, además, debería extenderse a todas las evidencias obtenidas mediante el uso de métodos cualitativos, entre las que se incluyen la obtenidas durante el desarrollo y la adaptación de un test y, en especial, la recopilación de evidencia relacionada con el contenido del test y también la relativa a puntos de corte para la interpretación de puntuaciones referida al criterio. En nuestra opinión, como mínimo, sería necesario especificar por separado cuántas personas participaron y su cualificación para ello, las estrategias que se utilizaron para obtener la información, los datos que se obtuvieron, sean verbales o numéricos, y las conclusiones y decisiones que de ellos se derivaron, para así facilitar la formación de un juicio propio de quien lee sobre la calidad de dichas decisiones.

Finalmente, respecto a la fiabilidad, los estudios de consistencia interna incluidos en los manuales que se han evaluado reciben mayoritariamente una valoración entre buena y excelente, lo que significa que se publican coeficientes de fiabilidad elevados obtenidos en muestras suficientemente grandes. Sin embargo, hay que remarcar la gran cantidad de casillas sin información que se observan en este apartado de la Tabla 3. Aunque no todos los diseños para estudiar fiabilidad son aplicables a todos los test, tanto en el CET-R (apartado 2.12.1) como en anteriores revisiones (e.g., Hernández et al., 2015) se sugiere como buena práctica la posibilidad de aportar varios coeficientes de fiabilidad para cada escala o subescala y subpoblación. Por nuestra parte, creemos que ello podría concretarse en que cada test aporte datos de fiabilidad obtenidos al menos con dos tipos de diseños de los que se contemplan en el CET-R o, alternativamente, explicaciones claras sobre por qué no son objeto de preocupación otras fuentes de variación aleatoria más allá de las derivadas de la coherencia entre los ítems del test. Incluso sin utilizar otro diseño que el de consistencia interna, sería muy oportuno reconocer que no todas las puntuaciones de un mismo test tienen la misma fiabilidad, lo que conllevaría la publicación de datos de fiabilidad relativa como los que se pueden obtener utilizando la teoría de respuesta a los ítems. Con esta sugerencia nos sumamos de nuevo a las recomendaciones realizadas en anteriores revisiones (e.g., Gómez, 2019).

Aportaciones Novedosas de la Universidad al Proceso de Revisión de Test

En esta edición queremos sumarnos al agradecimiento que, año tras año, se hace de la colaboración de las casas editoriales ofreciendo su ayuda y su experiencia durante el proceso de evaluación de test y del apoyo de la Comisión de Test del COP durante todo el proceso. Con mayor motivo si cabe porque nuestro proyecto docente ha conllevado mayor apoyo económico y ajuste a los tiempos académicos por su parte. Queremos también reflexionar sobre el proceso de priorización de los test a evaluar a partir de tres datos. El primer dato es la lista ordenada de los 25 test más usados por las personas colegiadas según una encuesta reciente (Muñiz et al., 2020). El segundo dato es la lista de los 84 test con informes publicados en la página web del COP desde el año 2010. Al comparar ambas listas se observa que solo coinciden en 12 casos y no siempre se trata de los test más usados. El tercer dato es la lista de los 91 test que han valorado en este mismo periodo nuestros estudiantes en sus trabajos de curso con mayor o menor éxito, una lista accesible contactando con la autora de correspondencia. Son test seleccionados entre los disponibles en nuestra docimoteca que, por su parte, realiza las adquisiciones atendiendo a demandas del profesorado de la Facultad de Psicología. Más de una cuarta parte de estos 91 coinciden con los test evaluados por el COP, pero resulta interesante observar la presencia de nueve test que están también en la lista de los más utilizados que no cuentan con informe oficial publicado. Otro dato interesante es que uno de los test que está en

la lista de los más utilizados y que no ha sido evaluado hasta ahora es un test no comercializado. Todo ello induce a pensar que, si se abriera la participación a nuevos colectivos con intereses en este proceso, probablemente se conseguiría mayor cobertura de las necesidades reales de información por parte de la profesión y, además, sería un nuevo paso en favor de la diversificación de las voces que participan en el proceso de revisión.

Respecto a la actuación de nuestro estudiantado, en la última columna de la Tabla 3 se presenta la comparativa de las puntuaciones promedio de la octava edición con las puntuaciones promedio del conjunto de las siete ediciones anteriores. Solo se incluyen las 10 características para las que ha sido posible obtener datos promedio en todas las revisiones disponibles. Aunque al observar dato por dato se encuentran algunas divergencias que se equilibran por ser unas al alza y otras a la baja, las dos series de datos tienen un mismo valor promedio de 3,9 puntos y el coeficiente de correlación entre ellas es de 0,90; es decir, se trata de evaluaciones globalmente muy comparables. Sea como fuere, no hay que perder de vista que las diferencias podrían atribuirse a quien evalúa, pero también a los test que se han sometido a evaluación en la presente edición.

Más informativa resulta la comparación de los informes de la Revisión 1 (especialista) con los de la Revisión 2 (estudiantes) sobre el mismo test. La mediana de los coeficientes de correlación entre las puntuaciones otorgadas por el Revisor 1 y el Revisor 2 en las preguntas en que todos los revisores dieron puntuaciones válidas a todos los test evaluados fue de 0,67, un valor moderado y muy similar al publicado por Ponsoda y Hontangas (2013) que fue de 0,61. De hecho, un valor como este o incluso menor sería de esperar en un proceso de revisión por pares en el que las discrepancias son consustanciales, tal como señalaron Fonseca-Pedrero y Muñiz (2017).

En las preguntas abiertas, los equipos de estudiantes ganadores escribieron textos mucho más largos (entre 2000 y 4400 palabras) que las personas profesionales (entre 800 y 3000 palabras). La lectura comparativa mostró que los textos del estudiantado pueden ser más argumentados, pero también más redundantes y más dependientes de la forma de presentación de la información utilizada en los libros de texto y en los manuales de los test. El profesorado atribuimos estos resultados a diversos motivos. Por una parte, puede ser que el estudiantado sienta la necesidad de escribir un apoyo teórico que le dé pie a expresar su opinión. Por otra, su formación previa está muy basada en la lectura de manuales educativos y de artículos científicos, por lo que podría resultarles difícil enfrentarse a textos desarrollados para la comercialización de un producto, aunque este tenga base científica. En cambio, las personas expertas tendrían más recursos para interpretar y valorar estos materiales. Otra explicación puede provenir de los comentarios del profesorado a los borradores, ya que en ellos se les pidió fundamentalmente que reconsiderasen las inconsistencias de sus valoraciones, que incorporasen información, y/o que desarrollasen más sus argumentos.

Relacionado con lo anterior, los equipos ganadores presentaron comentarios muy ajustados a las instrucciones y rúbricas del CET-R. Ello no es de extrañar porque, tal como hemos dicho, nuestras clases de psicometría están alineadas con ellas y las dudas del estudiantado al respecto se resolvieron gracias a la tutorización presencial. Con todo, las personas especialistas también han hecho comentarios en el sentido que algunas preguntas del CET-R son difíciles de contestar. Por ello, creemos que sería de gran ayuda que se implementara la propuesta de Fonseca-Pedrero y Muñiz (2017) de rebajar esta barrera facilitando más información técnica y desarrollando tutoriales sobre como rellenar el CET-R.

Todavía en el mismo orden de cosas, hemos detectado un amplio margen para seguir desarrollando lenguaje compartido tanto técnico como inclusivo. En cuanto al lenguaje técnico, nos alineamos con la opinión expresada en anteriores trabajos de que el CET-R es una buena guía para la construcción, la edición y el uso de los test (Elosua y Geisinger, 2016; Muñiz y Fonseca-Pedrero, 2019) y, por ello, sugerimos a las casas editoriales que al redactar los manuales se utilice en la medida de lo posible el lenguaje psicométrico tal como está expresado en el CET-R. Nada más lejos de nuestra intención que coartar la presentación de pruebas novedosas en apoyo de los usos de un test; antes al contrario, estas serían muy bienvenidas. Pero para la presentación de pruebas más clásicas, proponemos como referencia el lenguaje psicométrico del CET-R porque éste es una síntesis consensuada de algunos de los textos normativos más ampliamente aceptados como son los criterios de la EFPA, los estándares de la AERA y de la APA, de la *International Test Commission* y la norma ISO-10667 (Hernández et al., 2016) y también porque la homogeneización facilitaría enormemente el compartir el material con el resto de la profesión y, especialmente, con las personas principiantes. Y en relación con el uso de lenguaje inclusivo, por una parte, en los manuales de los test que hemos evaluado se utiliza ampliamente el masculino genérico para referirse a las personas y, por otra, en nuestra universidad se considera buena práctica valorar positivamente el uso de lenguaje inclusivo. Paradójicamente, nuestro estudiantado ha visto penalizada en sus escritos una práctica lingüística utilizada en los manuales que estaba evaluando. Creemos que es un buen momento para proponer a las casas editoriales que se sumen a dar ejemplo alineándose así con la política del Colegio Oficial de la Psicología en cuanto al uso de lenguaje inclusivo.

Finalmente, en la vertiente educativa, es de destacar que la asignatura ha tenido este curso una retención superior al 99% y un éxito del 93%, resultados dentro del rango de los obtenidos en los últimos cursos (UAB, n.d.-a) y que consideramos muy satisfactorios. Los pocos estudiantes que participaron en la encuesta de satisfacción reflejaron opiniones polarizadas. Entre las negativas destacó que el tiempo dedicado a elaborar el trabajo es mucho y compite con el tiempo dedicado a la explicación y asimilación de conceptos teóricos. Entre las positivas, predominó que el proyecto implica el fortalecimiento del aprendizaje conceptual mediante la apli-

cación real de la teoría y la aproximación al mundo profesional-laboral. Esta polarización se reflejó prácticamente en los mismos términos en las valoraciones del profesorado.

CONCLUSIONES Y RECOMENDACIONES

El primer objetivo de este trabajo ha sido valorar la calidad de 11 test sometidos a la octava edición de la evaluación de test en España aplicando el modelo de evaluación consensuado por la Comisión de Test del COP y reflejado en el cuestionario CET-R. Nuestras conclusiones han sido que en la documentación que acompaña a estos test se presentan datos más que adecuados en apoyo del desarrollo y adaptación del test así como sobre las muestras de tipificación. También se aportan excelentes datos relativos a la consistencia interna del test, pero debería incrementarse la presencia de otras pruebas de fiabilidad. Hemos detectado un mayor margen de mejora en cuanto a la aportación de pruebas en sustento de la validez del test para todos y cada uno de los usos propuestos. Nuestras propuestas de mejora son la asociación explícita de las pruebas de validez con cada uno de los usos propuestos, la especificación previa de las hipótesis que se pretende probar, el desarrollo de pruebas en favor de la equidad en el tratamiento de diversas personas evaluables, y el reporte de la metodología cualitativa basado en estándares actualizados.

Estas conclusiones tienen derivadas que afectan a la estructura y valoración de las preguntas del CET-R. En este sentido, recomendamos (a) incorporar una pregunta sobre los usos previstos del test en el apartado de descripción y estructurar la valoración de la validez en función de dichos usos, (b) dar dos opciones evaluativas para caracterizar la información faltante, que se calificaría o bien como *No relevante* o bien como *Relevante, pero no se aporta información*, (c) valorar de manera estructurada las evidencias de validez obtenidas con metodología cualitativa, considerando por separado el método y los resultados y (d) dar entidad a la aplicación equitativa de los test desarrollando de forma estructurada la valoración de los datos presentados sobre acomodaciones.

Nuestro segundo objetivo ha sido valorar la aportación novedosa de la universidad a dos aspectos del proceso de evaluación. En cuanto a los test que se someten a evaluación, nuestra conclusión ha sido que se podría ampliar la cobertura de las necesidades de información de la profesión si en el proceso de priorización se incorporara la opinión de otras entidades además de la Comisión de Test del COP. Por otra parte, hemos ofrecido datos sobre la ampliación de las voces representadas en la revisión ofreciendo esta oportunidad a estudiantes bajo la tutorización del profesorado. Nuestras conclusiones han sido que a nivel cuantitativo se ha mantenido una gran similitud con las ediciones anteriores y, a nivel narrativo, nuestros estudiantes han escrito textos más largos y más ajustados a las instrucciones del CET-R, aunque no siempre han proporcionado las argumentaciones más sólidas porque sus textos son muy dependientes del manual de la asignatura y de las valoraciones expresadas en la propia documentación que están evaluando. Ello nos ha llevado a destacar la conveniencia de profundizar en el desarrollo de lenguaje compartido entre los diferentes grupos con intereses en el uso de los test. En cuanto a la función educativa, concluimos que ha sido muy efectiva, puesto que nuestro estudiantado ha tenido éxito en la asignatura en un 93% de los casos, aunque también hemos observado opiniones polarizadas entre lo costoso y lo motivador del proyecto.

Nuestra conclusión sobre esta experiencia es que la incorporación de estudiantes bajo tutoría en el proceso de revisión de test ha resultado costosa en cuanto a los materiales y el tiempo necesarios para desarrollarla; variable en cuanto a su potencial motivador; y satisfactoria en relación con el éxito académico del estudiantado, la profesionalidad de los informes que han desarrollado, y las ideas que han aportado para desarrollar manuales de test que resulten cómodos para profesionales principiantes.

CONFLICTO DE INTERESES

No existe conflicto de intereses

AGRADECIMIENTOS

Agradecemos a la Dra. Laura Gómez y al Prof. Dr. José Muñiz su apoyo a lo largo de todo el proceso. A las casas editoriales, la aportación de seis ejemplares de cada test, a la Facultad de Psicología y a la Biblioteca de Humanidades de la UAB, la compra de los ejemplares faltantes hasta completar el material necesario para cubrir todos los grupos de clase. Al *Espai de Suport i Innovació Docent* de la Facultad de Psicología de la UAB su apoyo en la gestión del acceso a los materiales. A Remei Prat, Elena Ripollés, Salomé Tàrrega y Joan Pons su contribución al desarrollo del proyecto formando parte del equipo docente en cursos anteriores.

REFERENCIAS

AENOR. (2013). *Prestación de servicios de evaluación. Procedimientos y métodos para la evaluación de personas en entornos laborales y organizacionales. Parte 2: Deberes del proveedor de servicios*. (UNE-ISO 10667-2:2013) https://www.aenor.com/normas-y-libros/buscador-de-normas/UNE?c=N0051261

AERA, *American Psychological Association* [APA], y *National Council on Measurement in Education* [NCME]. (2014). *Standards for Educational and Psychological Testing* [*Estándares para los test educativos y psicológicos*]. American Educational Research Association.

COP. (2015a). *Directrices internacionales para el uso de los tests*. https://www.cop.es/index.php?page=directrices-internacionales

COP. (2015b). *Principios éticos de la evaluación en psicología*. https://www.cop.es/index.php?page=principios-eticos

Doval, E., Viladrich, C., Aliaga, J., Espelt, A., García-Rueda, R., Penelo, E., Prat, R. y Tárrega, S. (2013, 3–6 de septiembre). *Las asignaturas de contenido psicométrico en la UAB: saber y oficio* [comunicación]. XIII Congreso de la Asociación Española de Metodología de las Ciencias del Comportamiento. La Laguna, España.

Elosua, P., y Geisinger, K. F. (2016). Cuarta evaluación de tests editados en España: Forma y fondo. *Papeles del Psicólogo, 37*(2), 82–88.

Espelt, A., Viladrich, C., Doval, E., Penelo, E., y Aliaga, J. (2016, 5-7 de julio*). Relació entre l'adherència al funcionament de l'assignatura de Psicometria i la qualificació final dels estudiants [Relación entre la adherencia al funcionamiento de la asignatura de Psicometría y la calificación final de los estudiantes]* [poster]. IX Congreso Internacional de Docencia e Innovación Universitaria (CIDUI). Barcelona, España.

Evers, A., Muñiz, J., Hagemeister, C., Hstmælingen, A., Lindley, P., Sjöberg, A., y Bartram, D. (2013). Evaluación de la calidad de los tests: revisión del modelo de evaluación de la EFPA. *Psicothema, 25*(3), 283–291. https://doi.org/10.7334/psicothema2013.97

Fonseca-Pedrero, E., y Muñiz, J. (2017). Quinta evaluación de Tests editados en España: mirando hacia atrás, construyendo el futuro. *Papeles del Psicólogo, 38*(3), 161–168. https://doi.org/10.23923/pap.psicol2017.2844

Gómez, L. E. (2019). Séptima evaluación de test editados en España. *Papeles del Psicólogo, 40*(3), 205–210. https://doi.org/10.23923/pap.psicol2019.2909

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G., y Elosua, P. (2016). Cuestionario de Evaluación de Tests Revisado CET-R. *Papeles del Psicólogo, 37*, 161–168.

Hernández, A., Tomás, I., Ferreres, A., y Lloret, S. (2015). Evaluación de tests editados en España. *Papeles del Psicólogo, 36*(1), 1–8.

Hidalgo, M. D., y Hernández, A. (2019). Sexta evaluación de tests editados en España: Resultados e impacto del modelo en docentes y editoriales. *Papeles del Psicólogo, 40*(1), 21–30. https://doi.org/10.23923/pap.psicol2019.2886

International Test Commission. (2018). *ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations* [*Directrices de la ITC para la evaluación a gran escala de poblaciones lingüística y culturalmente diversas*]. https://www.intestcom.org/page/31

Levitt, H.M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R. y Suárez-Orozco, C. (2018). Journal Article Reporting Standards for Qualitative Primary, Qualitative Meta-Analytic, and Mixed Methods Research in Psychology: The APA Publications and Communications Board Task Force Report [Estándares de reporte en artículos de revistas para la investigación cualitativa primaria, meta-analítica cualitativa y de métodos mixtos en psicología: Informe del grupo de trabajo de la Junta de Publicaciones y Comunicaciones de la APA]. *American Psychologist, 73*(1). 26–46. https://dx.doi.org/10.1037/amp0000151

Muñiz, J., y Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test. [Ten steps for test development]. *Psicothema, 31*, 7–16. https://doi.org/10.7334/psicothema2018.291

Muñiz, J., Hernández, A., y Fernández-Hermida, J. R. (2020). Utilización de los test en España: el punto de vista de los psicólogos. *Papeles del Psicólogo, 41*(1). 1–15 https://doi.org/10.23923/pap.psicol2020.2921

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

Ponsoda, V., y Hontangas, P. (2013). Segunda evaluación de tests editados en España. *Papeles del Psicólogo, 34*(2), 82–90.

Redondo-Corcobado, P., y Fuentes, J. L. (2020). La investigación sobre el Aprendizaje-Servicio en la producción científica española: una revisión sistemática. *Revista Complutense de Educacion, 31*(1), 69–83. https://doi.org/10.5209/rced.61836_

UAB. (n.d.-a). *Seguiment de titulacions: Grau en Psicologia, Psicometria. [Seguimiento de titulaciones: Grado en Psicología, Psicometría]* http://siq.uab.cat/siq_public/titulacio-/2502443/assignatura/102569

UAB. (n.d.-b). *Enquesta de satisfacció d'assignatura de la UAB [Encuesta de satisfacción de assignatura de la UAB].* https://www.uab.cat/doc/QuestionariEnquestaAssignatures

Vermeulen, K. (2019). English version of the COTAN review system [Versión inglesa del sistema de revision COTAN] *Testing International, 41*, 8.

Viladrich, C., Doval, E., y Penelo, E. (2014, 23-25 de julio). *Student versus expert test reviews: What can we learn from them? [Revisiones de test de estudiantes versus especialistas: ¿qué podemos aprender de ellos?]* [comunicación en el simposio Viladrich, C. (presidencia) Symposium Tests review as a tool to enhancing testing practices]. VI European Congress of Methodology. Utrecht, Holanda.

Viladrich, C., Doval, E., Penelo, E., Aliaga, E., Espelt, A., García-Rueda, R., y Angulo-Brunet, A. (2019). *Avaluació de tests psicológics* [*Evaluación de test psicológicos*]. Assignatures i pràctiques ApS. http://pagines.uab.cat/aps/ca/content/assignatures-i-pràctiques-aps

Viladrich, C., Doval, E., Penelo, E., Aliaga, E., Espelt, A., García-Rueda, R., y Angulo-Brunet, A. (2021, 21-23 de julio). *Eighth edition of the Spanish evaluation of test quality: A service-learning experience [Octava evaluación de test editados en España: Una experiencia de aprendizaje-servicio].* [resumen aceptado en el simposio Hernández, A. y Muñiz, J. (presidencia) Improving tests and testing practices: international and multi-stakeholder perspectives]. IX European Congress of Methodology. Valencia, España.

Viladrich, C., Doval, E., Aliaga, J., Espelt, A., García-Rueda, R., Penelo, E., Tárrega, S., Ripollès, E., y Prat, R. (2014). Aprendices de certificadores en psicometría: una experiencia ABP con grupos grandes. *Revista del CIDUI, 2*,1–9.

Ziegler, M. (2014). Stop and state your intentions! Let's not forget the ABC of test construction [¡Detente y declara tus intenciones! No olvidemos el ABC de la construcción de test]. *European Journal of Psychological Assessment, 30*(4), 239–242. https://doi.org/10.1027/1015-5759/a000228