Papeles del Psicólogo, 2011. Vol. 32(2), pp. 113-128
http://www.cop.es/papeles

EVALUACIÓN DE TESTS EDITADOS EN ESPAÑA

José Muñiz[1], José R. Fernández-Hermida[1], Eduardo Fonseca-Pedrero[2], Ángela Campillo-Álvarez[1] y Elsa Peña-Suárez[1]
[1]*Universidad de Oviedo.* [2]*Universidad de La Rioja*

La utilización correcta de los tests psicológicos requiere por un lado que los instrumentos de medida tengan las propiedades psicométricas adecuadas, tales como fiabilidad y validez, y por otro, que los profesionales que los utilizan tengan la preparación técnica necesaria para usarlos. En el presente trabajo se presentan las primeras evaluaciones de tests editados en España, llevadas a cabo con un Modelo de Evaluación desarrollado por la Comisión Europea de Tests y adaptado al contexto español. El modelo permite llevar a cabo una evaluación tanto cualitativa como cuantitativa de las pruebas. Se evaluaron diez tests elegidos de entre los más utilizados por los profesionales españoles, cada uno de ellos se envió a dos revisores expertos para su evaluación, y a partir de dichos informes se elaboró el informe final. En líneas generales puede afirmarse que la calidad de los diez instrumentos de medida evaluados es buena, poniéndose de manifiesto sus puntos fuertes y débiles. A la vista de las revisiones se recomienda una mejora de las pruebas y sus Manuales para futuras ediciones, haciendo hincapié en la necesidad de incluir el mayor número posible de evidencias de validez sobre las pruebas. Finalmente se comentan los detalles del proceso de revisión seguido, y se analizan las posibles líneas de futuro en la evaluación de los tests en España.
***Palabras clave**. Tests, Uso de los tests, Evaluación de tests, Psicometría.*

The proper use of psychological tests requires that the measuring instruments have adequate psychometric properties, such as reliability and validity, and professionals who use those instruments have the necessary expertise to utilize them. In this paper we present the first evaluation of tests published in Spain, carried out with an Assessment Model developed by the European Test Commission, and adapted to the Spanish context. The model allows conducting a qualitative and quantitative evaluation of the test. Ten tests were evaluated, elected from among the most used by Spanish professionals. Each test was sent to two peer reviewers for evaluation, based on these reports a final inform was prepared. In light of the revisions carried out some improvements are suggested for future editions of the tests, emphasizing the need to include in the Manuals as many as possible evidences of validity of the tests. Finally, we discuss the details of the review process followed, and analyze possible future directions for the evaluation of tests in Spain.
***Key words:** Tests, Test use, Test evaluation, Psychometrics.*

La utilización correcta de los instrumentos de medida en cualquier campo profesional, y la psicología no es una excepción, requiere por un lado que los instrumentos tengan las propiedades métricas adecuadas, tales como fiabilidad y validez, y por otro, que los profesionales que los utilizan tengan la preparación técnica necesaria para usarlos. Un buen instrumento desde un punto de vista psicométrico puede echarse a perder si quien lo utiliza no tiene las competencias necesarias para su uso, bien podría decirse aquí, parafraseando al clásico, que a veces parece destino de los mejores tests caer en manos de los peores usuarios. Los colegios profesionales y distintas organizaciones nacionales e internacionales vienen haciendo esfuerzos desde hace bastantes años para intentar mejorar estos dos aspectos, la calidad de los tests y la preparación de los profesionales. Una exposición de estas actividades y proyectos puede consultarse en los trabajos de Muñiz y Bartram (2007) o Muñiz y Fernández-Hermida (2010). Naturalmente garantizar el correcto uso de los tests es condición necesaria, pero no suficiente, para que todo el proceso de evaluación psicológica llegue a buen término (Fernández-Ballesteros, De Bruyn, Godoy, Hornke, Ter Laak y Vizcarro, 2001).

Una de las demandas más reclamadas por los psicólogos profesionales cuando expresan sus opiniones sobre el uso de las pruebas es la necesidad de disponer de información técnica sobre los tests, que les ayude a tomar las decisiones adecuadas (Evers, Muñiz, Bartram et al., en prensa; Muñiz y Fernández-Hermida, 2000, 2010; Muñiz et al., 1999, 2001). Para dar respuesta a esta demanda de los psicólogos europeos, la Comisión de Tests de la Federación Europea de Asociaciones de Psicólogos (EFPA-SCTT), desarrolló un modelo de evaluación de

Correspondencia: José Muñiz. *Facultad de Psicología. Universidad de Oviedo. Plaza Feijoo, s/n. 33003 Oviedo. España.*
E-mail: jmuniz@uniovi.es

tests que puso a disposición de los profesionales de los países europeos, pudiendo consultarse en la página web: (http://www.efpa.eu/professional-development/tests-and-testing). En España el modelo fue adaptado por Prieto y Muñiz (2000) y publicado en esta misma revista, véase el Apéndice 1. La característica fundamental de este modelo europeo respecto a otros previos, como el desarrollado en Inglaterra (Bartram, 1996, 1998), o en Holanda (Evers, 2001a, 2001b), es que permite una evaluación exhaustiva de las distintas propiedades psicométricas de los tests, y además ofrece una evaluación tanto cuantitativa como cualitativa de la prueba. Este modelo se viene utilizando en varios países europeos para la evaluación de los tests, destacando Inglaterra y Holanda, en este último país todos los tests editados han sido evaluados con este modelo, o modelos previos (Evers et al., 2010).

EVALUACIÓN DE TESTS EN ESPAÑA

Como acabamos de señalar más arriba, en España se adaptó el modelo de evaluación de tests europeo (Prieto y Muñiz, 2000), pero no se había utilizado de forma sistemática hasta la fecha. En el año 2010 la Comisión de Tests del COP decidió por unanimidad iniciar el proceso de evaluación de tests. Para iniciar el proceso se eligieron diez pruebas, teniendo en cuenta tanto su uso por parte de los psicólogos españoles (Muñiz y Fernández-Hermida, 2000, 2010), como el interés de los editores para someter sus tests a esta primera evaluación. Siguiendo esos dos criterios, se sometieron a evaluación los diez tests que aparecen en la Tabla 1.

TABLA 1
RELACIÓN DE TESTS EVALUADOS

Test evaluados	
WAIS-III	Escala de Inteligencia de Wechsler para adultos - III
WISC-IV	Escala de inteligencia de Wechsler para niños-IV
MCMI-III	Inventario Clínico Multiaxial de Millon-III
MMPI-2-RF	Inventario Multifásico de la Personalidad de Minnesota-2 Reestructurado
16PF-5	Dieciséis Factores de Personalidad, quinta edición
PROLEC-R	Batería de Evaluación de procesos Lectores, revisada
EFAI	Evaluación Factorial de la Aptitudes Intelectuales
NEO PI-R	Inventario de Personalidad NEO Revisado
EVALUA	Batería Psicopedagógica
IGF	Batería de Inteligencia General y Factorial

PROCESO DE EVALUACIÓN

Una vez seleccionados los diez tests que se iban a revisar se siguió un proceso de evaluación por pares, similar al que se utiliza para revisar los artículos y proyectos de investigación científicos. La Comisión de Tests del COP seleccionó a un conjunto de revisores, y cada test se envió a dos de ellos. Se pretendió que el perfil de uno de ellos fuese más de carácter técnico-psicométrico y el otro más orientado a los aspectos sustantivos de la variable medida por el test. No se logró ese balance para todos los tests, pero sí en la mayoría de los casos, en la tabla 2 aparecen los veinte revisores que evaluaron los diez tests.

Los editores facilitaron gratuitamente dos juegos de cada test que fueron enviados a los revisores correspondientes, una vez finalizada la evaluación las pruebas fueron donadas a los revisores, además se les hizo un pago simbólico de cincuenta euros. La respuesta de los revisores a los que se enviaron los tests puede calificarse de excepcional, la tasa de rechazos de la invitación a revisar fue mínima, y siempre por razones de fuerza mayor. Desde aquí, y en nombre de la Comisión de Tests del COP, queremos expresar nuestro más sincero agradecimiento por su colaboración, nada de esto se hubiese podido hacer sin su ayuda. Una vez recibidas las eva-

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

Revisor	Afiliación
María Victoria del Barrio Gándara	UNED
Elisardo Becoña	Universidad de Santiago de Compostela
María José Blanca Mena	Universidad de Málaga
Isabel Calonge	Universidad Complutense de Madrid
Antonio Cano Vindel	Universidad Complutense de Madrid
Eduardo Fonseca Pedrero	Universidad de La Rioja
María Forns	Universidad de Barcelona
Jesús Enrique de la Fuente Arias	Universidad de Almería
Olaya García	Universidad de Barcelona
Juana Gómez Benito	Universidad de Barcelona
Héctor González Ordi	Universidad Complutense
María Dolores Hidalgo Montesinos	Universidad de Murcia
Serafín Lemos Giráldez	Universidad de Oviedo
José Antonio López Pina	Universidad de Murcia
Carmen Moreno	UNED
José Luis Miralles	Universidad de Valencia
María José Navas	UNED
José Carlos Núñez	Universidad de Oviedo
Vicente Ponsoda	Universidad Autónoma de Madrid
Celestino Rodríguez	Universidad de Oviedo

luaciones, el grupo de Psicometría de la Universidad de Oviedo, coordinado por José Muñiz, llevó a cabo la realización de un informe conjunto, en el cual se combinaban las evaluaciones de ambos revisores. Como ocurre con la evaluación de artículos científicos, o de proyectos de investigación, este informe no es la mera suma de los informes de los revisores, se tienen en cuenta sus valoraciones y se procede a generar un informe que refleje lo mejor posible sus opiniones. En ningún caso hubo que enviar el test a un tercer revisor, pues si bien en algunos casos había ciertas discrepancias y matices, pudieron ser resueltos de forma satisfactoria. Una vez elaborado este informe final se envió a los editores para que tanto ellos como los autores tuviesen la oportunidad de exponer su punto de vista. Las respuestas de los editores y autores fueron altamente profesionales, permitiendo aclarar algunos aspectos que no estaban suficientemente claros en los informes de los revisores. Este paso de dar la oportunidad de opinar a los autores y editores nos parece fundamental por dos razones, por un lado pueden ver de primera mano los puntos fuertes y débiles de su prueba, tomando conciencia en algunos casos de la necesidad de modificar algunos aspectos de la prueba en sucesivas versiones; y por otro, permite matizar información que puede habérseles pasado por alto a los revisores. Naturalmente, la opinión de los editores y autores no significa que se vaya a modificar el informe por el mero hecho de que eventualmente tengan algunas discrepancias, pero sí permite corregir y precisar algunas opiniones de los revisores. Nótese que el proceso de evaluación de los tests no constituye un ajuste de cuentas con los editores y autores, el objetivo fundamental es poner de manifiesto los puntos fuertes y débiles de las pruebas para así colaborar a que mejoren en las sucesivas versiones que necesariamente vendrán. Los tests son instrumentos vivos, no están hechos de una vez para todas, la idea es que las sucesivas versiones vayan aportando evidencias de validez que hagan más consistente y rigurosa la prueba.

En la tabla 3 se puede ver el resumen de las evaluaciones de las pruebas, como se puede observar, en líneas generales estas pruebas tienen un nivel de calidad muy razonable, cada una de ellas con unos puntos fuertes y otros más débiles. Las evaluaciones completas pueden consultarse en la página web del COP: www.cop.es, sección Comisión de Tests.

ALGUNAS LECCIONES APRENDIDAS

Sin entrar en el detalle de cada prueba, sí que cabría hacer algunas recomendaciones generales. Una primera recomendación es la necesidad de mejorar los Manuales, pues constituyen la piedra angular sobre la que se asienta la aportación de evidencias de validez de las pruebas. Los Manuales que se necesitan actualmente es-

TABLA 3
RESUMEN DE LAS CALIFICACIONES DE LOS TESTS EVALUADOS

	Test							
Características	**WISC-IV**	**EVALUA**	**MMPI-2-RF**	**16PF**	**PROLEC-R**	**EFAI**	**NEO PI-R**	**IGF**
Calidad de los Materiales y documentación	5	3,5	5	4,5	5	4,5	4	3
Fundamentación teórica	4,5	2,5	5	4,5	5	4	3,5	3,5
Adaptación Española	4,5	-	4	3	-	-	5	-
Análisis de los ítems	5	3,5	-	3	-	4,5	3,5	2
Validez de contenido	5	3	4,5	4	5	4	4	3,5
Validez de constructo	3	4	4,5	4	4	4	3	3
Análisis del sesgo	-	-	-	-	-	-	-	-
Validez predictiva	4	-	4	-	3	4	-	3
Fiabilidad: equivalencia	-	-	-	-	-	-	-	4
Fiabilidad: consistencia interna	4	3	4,5	3,5	3	5	4,5	3
Fiabilidad: estabilidad	3,5	-	4		-	-	-	-
Baremos	4	4,5	4	4	3	5	4	3,5

Nota. Durante el periodo de evaluación de las pruebas, la situación comercial del WAIS-III y del MCMI-III ha cambiado, y dado que el editor no tiene garantizados los derechos de las pruebas ha solicitado que no se publicasen los informes.
Las puntuaciones de la tabla están hechas en una escala de 1 a 5, y corresponden a las siguientes valoraciones: 1 = inadecuada; 2 = adecuada pero con carencias; a partir de 2,5 = adecuada; a partir de 3,5 = buena; a partir de 4,5 = excelente. Cuando aparece un guión (-) significa que no se aporta información o no procede.

tán lejos del tipo de folleto que se editaba antaño, en el que aparecían los baremos y poco más. Bien podría decirse que un test vale tanto como su Manual, el cual debe reflejar todas las evidencias y datos relativos al test, amén de una lista de referencias actualizadas sobre la prueba. Por ejemplo, un aspecto que se echa de menos en la mayoría de los Manuales analizados es un planteamiento explícito sobre la validez de contenido, es decir, explicar de qué manera se garantizó que la prueba contiene una representación adecuada de ítems para evaluar el constructo que sea. Esto no significa que las pruebas analizadas carezcan de validez de contenido, pero sí que no se ha hecho el esfuerzo de exponer clara y directamente la estrategia seguida para garantizar la validez de contenido, dándolo a veces por supuesto. Asimismo, en ninguno de los tests revisados se llevaron a cabo análisis sistemáticos sobre el Funcionamiento Diferencial de los Ítems, o sesgo, lo cual es un aspecto claramente mejorable, siendo de esperar que en futuras revisiones se vayan incorporando este tipo de análisis. Hay que asegurarse de que los ítems del test funcionan de forma similar para los distintos grupos implicados en la evaluación, tales como hombres y mujeres, distintas edades, o procedencia. En suma, los manuales, y por ende los tests, tienen que ir incorporando los nuevos avances psicométricos (Abad, Olea, Ponsoda y García, 2011; AERA, APA, NCME, 1999; Bartram y Hambleton, 2006; Bennett, 2006; Downing y Haladyna, 2006; Drasgow, Luecht y Bennett, 2006; Wilson, 2005).

Otro aspecto mejorable es que en el caso de pruebas adaptadas de otros países, mayormente Estados Unidos, no se incluyen de forma exhaustiva las evidencias de validez ya obtenidas en el país de origen. No es que esto exima de obtener evidencias en población española, pero ayuda a ir acumulando datos en torno a la prueba. Asimismo, en algunos Manuales no se especifica con detalle el proceso de traducción-adaptación de las pruebas, ni las equivalencias, si las hubiere, entre las formas originales y las adaptadas (Hambleton, Merenda, y Spielberger, 2005; Muñiz y Hambleton, 1996; van de Vijver y Hambleton, 1996).

En definitiva, un instrumento de medida permite a los profesionales llevar a cabo inferencias a partir de las puntuaciones obtenidas por las personas en la prueba, por tanto en los Manuales deben de aportarse con detalle y rigor las evidencias que garanticen que dichas inferencias se pueden hacer de forma fiable y válida. Es verdad que cada prueba tiene sus propias características y peculiaridades, pero en todos los casos se debe de informar a los profesionales de qué inferencias están documentadas y cuáles no.

PENSANDO EN FUTURAS EVALUACIONES

Lo que presentamos aquí es el inicio de la evaluación sistemática de tests en España, puede decirse que la experiencia es claramente positiva y el proceso debe de seguir a buen ritmo, la estación término sería que algún día se hubiesen evaluado todos los tests editados en nuestro país, como ocurre actualmente en Holanda. Comentamos a continuación algunas de las lecciones aprendidas en este primer intento, esperando que sean útiles para mejorar la práctica evaluadora futura. Un asunto que se presta a una cierta discusión es cuál es la mejor forma de elaborar el informe final del test a partir de las evaluaciones de los revisores. No existe una solución única e inapelable, tal como se hizo aquí funciona bien, estamos razonablemente satisfechos, pero hay otras posibles opciones. Se podría encargar el informe final a un guía experto en la prueba que combinaría los informes de los revisores y resolvería cualquier discrepancia, dado su conocimiento de la prueba. Esta es la aproximación que siguen los ingleses. Por su parte los holandeses, con casi treinta años de experiencia (Evers et al., 2010), hacen que los revisores interactúen hasta que logran un acuerdo sobre la prueba. Habrá que decidir cuál de los modelos se va a seguir en el futuro, incluso cabe una combinación de ellos.

Otro aspecto a dilucidar cara al futuro es qué tests elegir para continuar la evaluación, en el caso de estos diez primeros fue un acuerdo de la Comisión de Tests del COP, tal vez cara al futuro se pueda abrir la vía de que los editores por su parte sometan a valoración las pruebas que consideren oportunas, sin merma de que la propia Comisión de Tests elija por su parte pruebas que sean de gran interés para los profesionales.

En cuanto al Modelo Europeo de Evaluación de Tests, en el que se basa el utilizado aquí, (Apéndice 1), se encuentra actualmente en revisión por una comisión europea. Una vez que se disponga de la nueva versión se introducirán los cambios que procedan en nuestro modelo. Los puntos más calientes a revisar y sobre los que se está trabajando son la evaluación a distancia vía Internet, la realización de Informes Automatizados, la tecnología de la Teoría de Respuesta a los Ítems, los Tests Informatizados, y todo lo relativo a los Tests Referidos al Criterio. Es de interés ir incluyendo estos aspectos en el

modelo, sobre todo a medida que se vayan evaluando tests desarrollados con nuevas tecnologías psicométricas, si bien para la revisión de los tests más clásicos ello no supone ninguna merma, puesto que no suelen incluir estos aspectos, ahora bien a medida que se vaya ampliando la evaluación seguramente habrá que revisar tests que incluyan estas tecnologías, por lo que es conveniente que el modelo las contemple.

En relación a la aplicación del modelo en la práctica no se detectaron problemas graves por parte de los revisores, si bien se comentan a continuación algunos aspectos que se pueden mejorar. Así, por ejemplo, algunos revisores no llevan a cabo las valoraciones generales del test de forma cuantitativa tal y como se pide en la tabla diseñada a tal efecto. Se ve que las instrucciones no son lo suficientemente claras, por lo que sería conveniente especificar que las valoraciones se deben de hacer a partir del cálculo de la media aritmética de las puntuaciones otorgadas en los diferentes apartados que se especifican en la tabla de las valoraciones generales. Otras cuestiones se refieren a una confusión de conceptos, por ejemplo la administración oral se considera a veces un tipo de soporte (apartado 1.15), lo cual no es exacto. También genera cierta confusión determinar la existencia o no de diferentes formas del test (apartado 1.18). Algunos revisores incluyen en este apartado la posibilidad de obtención de informes informatizados, si bien el Modelo de Evaluación se refiere a si existen formas paralelas, versiones abreviadas, o versiones informatizadas.

Asimismo, sería conveniente reformular el apartado procedimiento de corrección (apartado 1.19), o explicar a qué hace referencia cada una de las categorías incluidas en el mismo, a veces *Lectora óptica* y *Automatizada por ordenador* tienden a confundirse. Los revisores suelen entender que la lectora óptica es un método de corrección automatizado a través del ordenador. En cuanto a contenidos que pueden incluirse en una futura versión del CET tienen que ver con la descripción general del test. Podría resultar interesante pedir que se detallen todas las revisiones que se han hecho desde la primera publicación del cuestionario que se evaluen (apartado 1.9) e insistir que es necesaria una descripción de las escalas que conforman el test revisado, ya que algunos revisores se limitan únicamente a su enumeración. Incluso en caso de existir varias subescalas o apartados sería conveniente que se especifiquen el número de ítems en cada una de ellas (apartado 1.14). Resultaría adecuado reformular de forma cuantitativa el ítem que evalúa la bibliografía básica aportada en el manual y que ha servido para elaboración del cuestionario. Podría preguntarse en estos términos: *Valora la bibliografía básica acerca del test a portada en la documentación como: inadecuada (1); adecuada pero con carencias (2); suficiente (3); buena (4); excelente (5).*

Otras sugerencias tienen que ver con la ampliación del apartado adaptación del test. Dada la relevancia e importancia del proceso de traducción y adaptación de pruebas extranjeras, que en el caso de muchas pruebas resulta deficiente, resultaría interesante obtener más información relativa a este proceso. Por ejemplo elaborar ítems relativos a métodos de adaptación como traducción inversa o doble traducción, qué tipo de profesionales han realizado tales tareas, si se han seguido las directrices internacionales pertinentes, etc. Estas y otras cuestiones más generales ya citadas deberán ser tenidas en cuenta a la hora de elaborar una nueva versión del Cuestionario de Evaluación de Tests.

AGRADECIMIENTOS

Deseamos expresar nuestro más sincero agradecimiento a los miembros de la Comisión de Tests del Colegio Oficial de Psicólogos, sin su ayuda y colaboración este trabajo no se hubiese podido realizar. Muchas gracias a Eduardo Montes Velasco, Rocío Fernández Ballesteros, Miguel Martínez García, Jaime Pereña Brand y Javier Rubio Ramiro.

REFERENCIAS

Abad, F. J., Olea, J., Ponsoda, V. & García, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Síntesis.

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

Bartram, D. (1996). Test qualifications and test use in the UK: The competence approach. *European Journal of Psychological Assessment, 12*, 62-71.

Bartram, D., & Hambleton, R. K. (Eds.) (2006). *Computer-based testing and the Internet*. Chichester, UK: Wiley and Sons.

Bennett, R. E. (2006). Inexorable and inevitable: The continuing story of technology and assessment. En D. Bartram & R. K. Hambleton (Eds.), *Computer-based*

testing and the Internet (pp. 201-217). Chichester, UK: Wiley and Sons.

Downing, S. M., & Haladyna, T. M. (Eds.) (2006). *Handbook of test development*. Hillsdale, NJ: Erlbaum.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. En R. L. Brennan (Ed.), *Educational measurement* (pp. 471-515).Westport, CT: ACE/Praeger.

Evers, A. (2001a). Improving test quality in the Netherlands: Results of 18 years of test ratings. *International Journal of Testing, 1*, 137-153.

Evers. A. (2001b). The revised Dutch rating system for test quality. *International Journal of Testing, 1*, 155-182.

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., Frans, O., Gintiliené, G., Hagemeister, C., Halama, P., Iliescu, D., Jaworowska, A., Jiménez, P., Manthouli, M., Matesic, K., Schittekatte, M., Sümer, C., & Urbánek, T. (en prensa). Testing practices in the 21st century: Developments and European psychologists's opinions. *European Psychologist*.

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.

Eyde, L. D., Robertson, G. J., Krug, S. E., Moreland, K. L., Robertson, A. G., & Shewan, C. M., et al. (1993). *Responsible test use. Case studies for assessing human behavior*. Washington, DC: American Psychological Association.

Fernández-Ballesteros, R., De Bruyn, E., Godoy, A., Hornke, L., Ter Laak, J., Vizcarro, C., et al. (2001). Guidelines for the assessment process (GAP): A proposal for discussion. *European Journal of Psychological Assessment, 17*, 187-200.

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

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

Muñiz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., Fernández-Hermida, J.R., & Zaal, J. (2001). Testing practices in European countries. *European Journal of Psychological Assessment, 17*, 201-211.

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

Muñiz, J., & 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., & Hambleton, R. K. (1996). Directrices para la traducción y adaptación de los tests. *Papeles del Psicólogo*, 66, 63-70.

Muñiz, J., Prieto, G., Almeida, L., & Bartram, D. (1999). Test use in Spain, Portugal and Latin American countries. *European Journal of Psychological Assessment, 15*, 151-157.

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

van de Vijver. F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89-99.

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

APÉNDICE 1
CUESTIONARIO UTILIZADO PARA LA EVALUACIÓN DE LOS TESTS (CET)

1. Descripción general del test[1]

1.1. Nombre del test:
1.2. Nombre del test en su versión original (si la versión española es una adaptación):
1.3. Autor/es del test original:
1.4. Autor/es de la adaptación española:
1.5. Editor del test en su versión original:
1.6. Editor de la adaptación española:
1.7. Fecha de publicación del test original:
1.8. Fecha de publicación del test en su adaptación española:
1.9. Fecha de la última revisión del test en su adaptación española:
1.10. Clasifique el área general de la o las variables que pretende medir el test[2]

( ) Inteligencia
( ) Aptitudes
( ) Habilidades y Rendimiento académico
( ) Psicomotricidad
( ) Neuropsicología
( ) Personalidad
( ) Motivación
( ) Actitudes
( ) Intereses
( ) Escalas de Desarrollo
( ) Competencia Curricular
( ) Escalas Clínicas
( ) Potencial de Aprendizaje
( ) Otros (Indique cuál:.....)

1.11. Breve descripción de la variable o variables que pretende medir el test:
(Se trata de hacer una descripción no evaluativa del test entre 200-600 palabras. La descripción debe de proporcionar al lector una idea clara del test, lo que pretende medir y las escalas que lo conforman)

1.12. Área de aplicación[3]

( ) Psicología clínica
( ) Psicología educativa
( ) Neuropsicología
( ) Psicología forense
( ) Psicología del trabajo y las organizaciones
( ) Psicología del deporte
( ) Servicios sociales
( ) Psicología del Tráfico
( ) Otros (Indique cuál:.....)

[1] Si el test está compuesto de subtests heterogéneos en su formato y características, rellene un cuestionario para cada subtest.
[2] Puede marcar más de una opción.
[3] Puede marcar más de una opción.

1.13. Formato de los ítems[4]:
( ) Respuesta libre
( ) Respuesta dicotómica (si/no, verdadero/falso, etc)
( ) Elección múltiple
( ) Tipo Likert
( ) Adjetivos bipolares
( ) Otro (Indique cuál:.....)

1.14. Número de ítems[5]:

1.15. Soporte[6]:
( ) Administración oral
( ) Papel y lápiz
( ) Manipulativo
( ) Informatizado
( ) Otro (Indique cuál:.....)

1.16. Cualificación requerida para el uso del test de acuerdo con la documentación aportada:
( ) Ninguna
( ) Entrenamiento y Acreditación especifica*
( ) Nivel A[7]
( ) Nivel B
( ) Nivel C
( ) Otra (Indique cuál:.....)

*Indique el nombre de la institución que lleva a cabo la acreditación:

1.17. Descripción de las poblaciones a las que el test es aplicable (especifique el rango de edad, nivel educativo, etc., y si el test es aplicable en ciertas poblaciones específicas: minorías étnicas, discapacitados, grupos clínicos, etc.):

1.18. Indique si existen diferentes formas del test y sus características (formas paralelas, versiones abreviadas, versiones informatizadas o impresas, etc). En el caso de que existan versiones informatizadas, describa los requisitos mínimos del hardware y software.

1.19. Procedimiento de corrección:
( ) Manual mediante plantilla
( ) Lectora óptica
( ) Automatizada por ordenador
( ) Efectuado exclusivamente por la empresa suministradora
( ) Mediante expertos
( ) Hoja Autocorregible
( ) Otro (Indique cuál:.....).

[4] Puede marcar más de una opción.
[5] Si el test tiene varias escalas, indique el número de ítems de cada una.
[6] Puede marcar más de una opción.
[7] Algunos países han adoptado sistemas para la clasificación de los tests en distintas categorías, en función de la cualificación requerida por los usuarios. Estos sistemas de clasificación proporcionan a los editores de tests un medio para decidir a quién pueden vender los tests. Un sistema muy utilizado es el que divide los tests en tres categorías: Nivel A (tests de rendimiento y conocimientos), Nivel B (tests colectivos de aptitudes e inteligencia) y Nivel C (tests de aplicación individual de inteligencia, personalidad y otros instrumentos complejos).

1.20. Puntuaciones: (Describa el procedimiento para obtener las puntuaciones directas).

1.21. Transformación de las puntuaciones:
( ) Característica no aplicable para este instrumento
( ) Normalizada
( ) No normalizada

1.22. Escalas utilizadas:
( ) Centiles
( ) Puntuaciones típicas
( ) Cocientes de desviación
( ) Eneatipos
( ) Decatipos
( ) T (Media 50 y desviación típica 10)
( ) S (Media 50 y desviación típica 20)
( ) Otra (Indique cuál:.....)

1.23. Posibilidad de obtener informes automatizados:
( ) No
( ) Si*

*En caso afirmativo haga una breve descripción no evaluativa del Informe Automatizado, en la que se hagan constar las características fundamentales, tales como tipo de informe, estructura, claridad, estilo, tono, etc.

1.24. El editor ofrece un servicio para la corrección y/o elaboración de informes:
( ) No
( ) Si

1.25. Tiempo estimado para la aplicación del test (instrucciones, ejemplos y respuestas a los ítems).
En aplicación individual:.....

En aplicación colectiva:.....

1.26. Documentación aportada por el editor:
( ) Manual
( ) Libros o artículos complementarios
( ) Disketes/CD
( ) Otra (Indique cuál:.....)

1.27. Precio de un juego completo de la prueba (documentación, test, plantillas de corrección; en el caso de tests informatizados no se incluye el costo del hardware):

1.28. Precio y número de ejemplares del paquete de cuadernillos (tests de papel y lápiz):

1.29. Precio y número de ejemplares del paquete de hojas de respuesta (tests de papel y lápiz):

1.30. Precio de la corrección y/o elaboración de informes por parte del editor:

1.31. Bibliografía básica acerca del test aportada en la documentación:

2. Valoración de las características del test

2.1. Calidad de los materiales del test (objetos, material impreso o software):

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (Impresión y presentación de gran calidad, software muy atractivo y eficiente, etc.)

2.2. Calidad de la documentación aportada:

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (Descripción muy clara y completa de las características técnicas, fundamentada en abundantes datos y referencias)

2.3. Fundamentación teórica:

( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (Descripción muy clara y documentada del constructo que se prete de medir y del procedimiento de medición)

2.4. Adaptación del test (si el test ha sido traducido y adaptado para su aplicación en España):

( ) Característica no aplicable para este instrumento
( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (Descripción precisa del procedimiento de traducción, de la adaptación de los ítems a la cultura española, de los estudios de equivalencia con la versión original, utilización de la normativa de la International Test Commission, etc.).

2.5. Calidad de las instrucciones:

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (Claras y precisas. Muy adecuadas para las poblaciones a las que va dirigido el test).

2.6. Facilidad para comprender la tarea:

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Suficiente
**** ( ) Buena
***** ( ) Excelente (Los sujetos de las poblaciones a las que va dirigido el test pueden comprender fácilmente la tarea a realizar).

2.7. Facilidad para registrar las respuestas:

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (El procedimiento para emitir o registrar las respuestas es muy simple por lo que se evitan los errores en la anotación).

2.8. Calidad de los ítems (aspectos formales):

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (La redacción y el diseño son muy apropiados)

2.9. Análisis de los ítems

2.9.1 Datos sobre el análisis de los ítems:

( ) Característica no aplicable para este instrumento
( ) No se aporta información en la documentación
* ( ) Inadecuados
** ( ) Adecuados pero con algunas carencias
*** ( ) Adecuados
**** ( ) Buenos
***** ( ) Excelentes (Información detallada sobre diversos estudios acerca de las características psicométricas de los ítems: dificultad o variabilidad, discriminación, validez, distractores, etc.)

2.10. Validez

2.10.1. Validez de contenido[8]:

2.10.1.1. Calidad de la representación del contenido o dominio:

***** ( ) Inadecuada
***** ( ) Adecuada pero con algunas carencias
***** ( ) Adecuada
***** ( ) Buena
***** ( ) Excelente (En la documentación se presenta una precisa definición del contenido. Los ítems muestrean adecuadamente todas las facetas del contenido)

2.10.1.2. Consultas a expertos[9]:

( ) No se aporta información en la documentación
* ( ) No se ha consultado a expertos sobre la representación del contenido
** ( ) Se ha consultado de manera informal a un pequeño número de expertos
*** ( ) Se ha consultado a un pequeño número de expertos mediante un procedimiento sistematizado (N<10)
**** ( ) Se ha consultado a un número moderado de expertos mediante un procedimiento sistematizado (10≤N≤30)
***** ( ) Se ha consultado a un amplio número de expertos mediante un un procedimiento sistematizado (N>30)

[8] Este aspecto es esencial en los tests referidos al criterio y particularmente en los tests de rendimiento académico. Emita su juicio sobre la calidad de la representación del contenido o dominio. Si en la documentación aportada aparecen las evaluaciones de los expertos, tómelas en consideración.

[9] Las cifras acerca del tamaño de las muestras y de los estadísticos que aparecerán más adelante tienen un carácter orientativo.

2.10.2. Validez de constructo:

2.10.2.1. Diseños empleados[10]:

( ) No se aporta información en la documentación
( ) Correlaciones con otros tests
( ) Diferencias entre grupos
( ) Matriz multirasgo-multimétodo
( ) Análisis factorial exploratorio
( ) Análisis factorial confirmatorio
( ) Diseños experimentales
( ) Otros (Indique cuales:.....).

2.10.2.2. Tamaño de las muestras en la validación de constructo:

( ) No se aporta información en la documentación
* ( ) Un estudio con una muestra pequeña (N<200)
** ( ) Un estudio con una muestra moderada (200≤N≤500)
*** ( ) Un estudio con una muestra grande (N>500)
**** ( ) Varios estudios con muestras de tamaño moderado
***** ( ) Varios estudios con muestras grandes

2.10.2.3. Procedimiento de selección de las muestras*:

( ) No se aporta información en la documentación
( ) Incidental
( ) Aleatorio
*Describa brevemente el procedimiento de selección.

2.10.2.4. Mediana de las correlaciones del test con otros tests similares:

( ) No se aporta información en la documentación
* ( ) Inadecuada (r<0.25)
** ( ) Adecuada pero con algunas carencias (0.25≤r<0.40)
*** ( ) Adecuada (0.40≤r<0.50)
**** ( ) Buena (0.50≤r<0.60)
***** ( ) Excelente (r≥0.60)

2.10.2.5. Calidad de los tests empleados como criterio o marcador:

( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente

2.10.2.6. Datos sobre el sesgo de los ítems:

( ) Característica no aplicable para este instrumento
( ) No se aporta información en la documentación
* ( ) Inadecuados
** ( ) Adecuados pero con algunas carencias
*** ( ) Adecuados

[10] Puede marcar más de una opción.

**** ( ) Buenos
***** ( ) Excelentes (Información detallada sobre diversos estudios acerca del sesgo de los ítems relacionado con el sexo, la lengua materna, etc. Empleo de la metodología apropiada)

2.10.3. Validez predictiva

2.10.3.1. Describa los criterios empleados y las características de las poblaciones:

2.10.3.1. Diseño de selección del criterio[11]:
( ) Concurrente
( ) Predictivo
( ) Retrospectivo

2.10.3.2. Tamaño de las muestras en la validación predictiva:
( ) No se aporta información en la documentación
* ( ) Un estudio con una muestra pequeña (N<100)
** ( ) Un estudio con una muestra moderada (100≤N<200)
*** ( ) Un estudio con una muestra grande y representativa (N≥200)
**** ( ) Varios estudios con muestras representativas de tamaño moderado
***** ( ) Varios estudios con muestras grandes y representativas

2.10.3.3. Procedimiento de selección de las muestras*:
( ) No se aporta información en la documentación
( ) Incidental
( ) Aleatorio

*Describa brevemente el procedimiento de selección.

2.10.3.4. Mediana de las correlaciones del test con los criterios:
( ) No se aporta información en la documentación
* ( ) Inadecuada (r<0.20)
** ( ) Suficiente (0.20≤r<0.35)
*** ( ) Buena (0.35≤r<0.45)
**** ( ) Muy buena (0.45≤r<0.55)
***** ( ) Excelente (r≥0.55)

2.10.4. Comentarios sobre la validez en general:

2.11. Fiabilidad

2.11.1. Datos aportados sobre la fiabilidad:
( ) Un único coeficiente de fiabilidad
( ) Un único error típico de medida
( ) Coeficientes de fiabilidad para diferentes grupos de sujetos
( ) Error típico de medida para diferentes grupos de sujetos

2.11.2. Equivalencia (Formas paralelas):

[11] Puede marcar más de una opción.

2.11.2.1. Tamaño de las muestras en los estudios de equivalencia:

( ) No se aporta información en la documentación
* ( ) Un estudio con una muestra pequeña (N<200)
** ( ) Un estudio con una muestra moderada (200≤N<500)
*** ( ) Un estudio con una muestra grande (N>500)
**** ( ) Varios estudios con muestras de tamaño moderado
***** ( ) Varios estudios con muestras grandes

2.11.2.2. Mediana de los coeficientes de equivalencia:

( ) No se aporta información en la documentación
* ( ) Inadecuada (r<0.50)
** ( ) Adecuada pero con algunas carencias (0.50≤r<0.60)
*** ( ) Adecuada (0.60≤r<0.70)
**** ( ) Buena (0.70≤r<0.80)
***** ( ) Excelente (r≥0.80)

2.11.3. Consistencia interna

2.11.3.1. Tamaño de las muestras en los estudios de consistencia:

( ) No se aporta información en la documentación
* ( ) Un estudio con una muestra pequeña (N<200)
** ( ) Un estudio con una muestra moderada (200≤N<500)
*** ( ) Un estudio con una muestra grande (N≥500)
**** ( ) Varios estudios con muestras de tamaño moderado
***** ( ) Varios estudios con muestras grandes

2.11.3.2. Mediana de los coeficientes de consistencia:

( ) No se aporta información en la documentación
* ( ) Inadecuada (r<0.60)
** ( ) Adecuada pero con algunas carencias (0.60≤r<0.70)
*** ( ) Adecuada (0.70≤r<0.80)
**** ( ) Buena (0.80≤r<0.85)
***** ( ) Excelente (r≥0.85)

2.11.4. Estabilidad (Test-Retest)

2.11.4.1. Tamaño de las muestras en los estudios de estabilidad[12]:

( ) No se aporta información en la documentación
* ( ) Un estudio con una muestra pequeña (N<100)
** ( ) Un estudio con una muestra moderada (100≤N<200)
*** ( ) Un estudio con una muestra grande (N≥200)
**** ( ) Varios estudios con muestras de tamaño moderado
***** ( ) Varios estudios con muestras grandes

[12] Número de sujetos con ambas puntuaciones (antes-después).

2.11.4.2. Mediana de los coeficientes de estabilidad:

( ) No se aporta información en la documentación
* ( ) Inadecuada (r<0.55)
** ( ) Adecuada pero con algunas carencias (0.55≤r<0.65)
*** ( ) Adecuada (0.65≤r<0.75)
**** ( ) Buena (0.75≤r<0.80)
***** ( ) Excelente (r≥0.80)

2.11. 5 Comentarios sobre la fiabilidad en general:

2.12. Baremos

2.12.1. Calidad de las normas:

( ) No se aporta información en la documentación
* ( ) Un baremo que no es aplicable a la población objetivo
** ( ) Un baremo aplicable a la población objetivo con cierta precaución
*** ( ) Un baremo adecuado para la población objetivo
**** ( ) Varios baremos dirigidos a diversos estratos poblacionales
***** ( ) Amplio rango de baremos en función de la edad, el sexo, el nivel cultural y otras características relevantes

2.12.2. Tamaño de las muestras[13]:

( ) No se aporta información en la documentación
* ( ) Pequeño (N<150)
** ( ) Suficiente (150≤N<300)
*** ( ) Moderado (300≤N<600)
**** ( ) Grande (600≤N<1000)
***** ( ) Muy grande (N≥1000)

2.12.3. Procedimiento de selección de las muestras*:

( ) No se aporta información en la documentación
( ) Incidental
( ) Aleatorio

*Describa brevemente el procedimiento de selección.

2.12.4. Comentarios sobre los baremos

3. Valoración global del test

3.1. Con una extensión máxima de 1000 palabras, exprese su valoración del test, resaltando sus puntos fuertes y débiles, así como recomendaciones acerca de su uso en diversas áreas profesionales. Indique asimismo cuáles son las características de la prueba que podrían ser mejoradas, carencias de información en la documentación, etc.

[13] Si hay varios baremos, clasifique el tamaño promedio

A modo de resumen, rellene las Tablas 1 y 2.

La Tabla 1 incluye algunos datos descriptivos del test.

TABLA 1
DESCRIPCIÓN DEL TEST

Característica		Descripción
Nombre del test	(apartado 1.1)	
Autor	(apartado 1.3)	
Autor de la adaptación española	(apartado 1.4)	
Fecha de la última revisión	(apartado 1.9)	
Constructo evaluado	(apartado 1.11)	
Areas de aplicación	(apartado 1.12)	
Soporte	(apartado 1.15)	

En la Tabla 2 se resume la valoración de las características generales del test. Tome en consideración el promedio de las calificaciones emitidas en los apartados que figuran en la segunda columna de la Tabla 2.

TABLA 2
VALORACIÓN DEL TEST

Característica	Apartados	Valoración
Materiales y documentación	2.1 y 2.2	
Fundamentación teórica	2.3	
Adaptación	2.4	
Análisis de ítems	2.9	
Validez de contenido	2.10.1	
Validez de constructo	2.10.2	
Análisis del sesgo	2.10.2.6	
Validez predictiva	2.10.3	
Fiabilidad: equivalencia	2.11.2	
Fiabilidad: consistencia interna	2.11.3	
Fiabilidad: estabilidad	2.11.4	
Baremos	2.12	