GRUPO EDITORIAL DE PSICOFUNDACIÓN

Papeles del Psicólogo (2023), 44(1), 1-7

Papeles del Psicólogo
Psychologist Papers
https://www.papelesdelpsicologo.es/ • ISSN 0214–7823

PAPELES DEL PSICÓLOGO
PSYCHOLOGIST PAPERS

Revista del Consejo General de la Psicología de España

Artículo

Novena evaluación de los test editados en España

Luis M. Lozano

Universidad de Granada, Centro de Investigación Mente Cerebro y Comportamiento (CIMCYC), España

INFORMACIÓN

Recibido: Noviembre 02, 2022
Aceptado: Diciembre 23, 2022

Palabras clave
Test
Evaluación de test
Psicometría
Propiedades psicométricas
CET-R

RESUMEN

La Comisión Nacional de Test, perteneciente al Consejo General de la Psicología de España, elabora anualmente una evaluación de los test editados en España. Para ello, en esta edición, se han involucrado tanto la Comisión como diferentes casas editoriales (TEA Hogrefe, Pearson Educación, GiuntiEOS y CEPE) y doce evaluadores (seis especialistas en la materia sustantiva y seis en Psicometría). La evaluación realizada se basa en el modelo europeo de evaluación de la calidad de los test adaptados al español, que ha dado como resultado el Cuestionario de Evaluación de Test Revisado (CET-R). Como resultado general cabe destacar que la calidad de los seis cuestionarios evaluados es buena y coherente con los resultados obtenidos en evaluaciones previas. Así mismo, se presentan diferentes aspectos de mejora en el proceso evaluativo.

Ninth review of tests published in Spain

Keywords
Tests
Assessing test quality
Psychometrics
Psychometric properties
CET-R

ABSTRACT

The National Test Commission, belonging to the Spanish Psychological Association, prepares an annual review of the tests published in Spain. In this edition, the Commission, different publishing houses (TEA Hogrefe, Pearson Education, GiuntiEOS, and CEPE), and twelve external evaluators (six specialists in the substantive subject and six experts in psychometrics) have been involved. The review carried out is based on the European model of evaluation of the quality of the tests adapted to Spanish, which has resulted in the Revised Test Evaluation Questionnaire (CET-R). As a general result, it should be noted that the quality of the six questionnaires evaluated is good and consistent with the results obtained in previous reviews. Furthermore, different aspects of improvement in the evaluation process are presented.

Cómo citar: Lozano, L. (2023). Novena evaluación de los test editados en España. *Papeles del Psicólogo, 44*(1), 1-7. https://doi.org/10.23923/pap.psicol.3004
Autor de correspondencia: lmlozano@ugr.es

La mayor parte de la información que obtienen y manejan los psicólogos para tomar cualquier tipo de decisión proviene de los cuestionarios. Esta afirmación puede considerarse transversal a las diferentes áreas de trabajo de la Psicología, ya que el uso de los cuestionarios psicométricos puede considerarse universal (Muñiz et al., 2020). Aunque, obviamente, la toma de decisiones le corresponde al profesional de la psicología y no a la prueba psicológica empleada, estas decisiones deben ser tomadas basándose en información de calidad. Siempre hay que tener presente que cualquier decisión que se tome afecta a la vida de las personas evaluadas, y por tanto, no puede ser tomada a la ligera.

Aunque, tal y como se comentó anteriormente, la calidad de la prueba empleada es crucial para una buena toma de decisiones, no es suficiente. No hay que restarles importancia a todas las partes implicadas en un proceso evaluativo. Por ello, aunque en este artículo se trate en mayor detalle la evaluación de la calidad de los test editados en España, también se pretende hacer una referencia, aunque somera, tanto al profesional de la psicología que realiza la evaluación como a la persona que es evaluada.

El profesional de la Psicología

En primer lugar, la formación de los profesionales de la psicología debe ser amplia, tanto en los diferentes aspectos sustantivos a evaluar como en lo psicométrico (Muñiz et al., 2011). Al igual que en la elaboración de un cuestionario las primeras preguntas que se deben realizar son ¿qué se quiere medir?, ¿a quién se pretende evaluar?, ¿para qué se quiere medir?, o ¿cómo se va a realizar la evaluación?, los profesionales de la Psicología deben plantearse estas mismas preguntas antes de realizar una evaluación. Las respuestas a estas preguntas le guiarán para seleccionar el cuestionario adecuado para los objetivos de la evaluación.

Ante la pregunta de *"¿qué se quiere medir?"* el psicólogo no se encuentra ante una respuesta sencilla. Si en algún ámbito de la Psicología se hace patente la falacia "jingle-jangle" es en este caso. La "falacia jingle" hace referencia a la utilización de un único término para describir constructos que son diferentes. Por su parte, la "falacia jangle" ocurre cuando se emplean diferentes términos para describir el mismo constructo.

Como es bien sabido, no existe una aproximación universal a los constructos, es decir, que diferentes teóricos al definir un constructo pueden seleccionar diferentes conductas para operacionalizarlo. Lógicamente, la selección de diferentes conductas lleva a que bajo una misma etiqueta (e.g., depresión) se encuentren cuestionarios que proceden de diferentes marcos teóricos, y que definen el constructo de manera diferente (falacia "jingle"). Por ello, el psicólogo debe ser capaz de seleccionar entre todos los cuestionarios aquellos que se encuadran dentro del marco teórico en el que se está moviendo. Pero a su vez, debe tener los conocimientos sustantivos suficientes como para poder determinar que un cuestionario que evalúa un constructo denominado de una manera diferente al que se pretende medir puede capturar las conductas en las que está interesado (falacia "jangle"; Gonzalez et al., 2021)

En relación con *"¿a quién se quiere evaluar?"* el profesional de la psicología debe ser conocedor de las posibles diferencias entre grupos en la variable a evaluar. De este modo, puede seleccionar el cuestionario que más se adapte a las características de las personas a evaluar. En primer lugar, la selección del grupo de referencia es de vital importancia. Comparar las puntuaciones de una persona con puntuaciones de personas que no son de su grupo normativo lleva, irremediablemente, a evaluar incorrectamente a las personas. Por otro lado, el psicólogo también debe conocer si existen acomodaciones o si el cuestionario ha sido elaborado siguiendo las recomendaciones de "diseño universal" (AERA et al., 2014).

La respuesta a *"¿para qué se quiere medir?"* también es de vital importancia. Existen múltiples cuestionarios que evalúan un constructo desde el mismo marco teórico, pero que han sido creados con finalidades diferentes. Hay que tener en cuenta que los ítems que componen un cuestionario no tienen por qué ser los mismos cuando la finalidad de un cuestionario es de screening poblacional o una evaluación clínica. Conviene recordar que las pruebas no son o dejan de ser válidas, las afirmaciones sobre la validez deben referirse a la interpretación que se hace de las puntuaciones para un uso determinado (AERA et al., 2014). Por ello, determinar la utilización que se hará de la puntuación y buscar cuestionarios que hayan mostrado evidencias de validez para dicho uso es una labor central del evaluador.

Para responder a *"¿cómo se va a realizar la evaluación?"* el psicólogo debe saber que existen múltiples procedimientos de aplicación de pruebas (e.g., papel y lápiz, informatizado, adaptativo…) que pueden y deben adaptarse a las características de las personas evaluadas. El procedimiento de evaluación debe depender de la experiencia que tienen las personas evaluadas en la respuesta a cuestionarios, la experiencia con ordenadores… No tener en cuenta estos condicionantes puede generar variables extrañas que afecten a la calidad de los datos obtenidos.

Aunque hasta el momento solo se está poniendo el énfasis en la relación entre el psicólogo y el cuestionario, existen otros aspectos que deben ser tenidos en cuenta. Los profesionales de la psicología deben saber realizar la evaluación de forma correcta, debe establecer un ambiente adecuado para la evaluación, deben interpretar de forma adecuada los resultados obtenidos en la prueba, deben ser capaces de transmitir la información obtenida de una forma clara, comprensible y útil (este último punto es frecuentemente olvidado), y están obligados a realizar un uso ético de las puntuaciones obtenidas.

Con la finalidad de mejorar la formación de los profesionales se han establecido diferentes líneas de trabajo desde distintas asociaciones. Para el lector interesado en las diferentes propuestas realizadas se recomienda el trabajo de Muñiz et al. (2020).

La persona evaluada

Aunque los profesionales de la psicología no pueden reglamentar las conductas de las personas evaluadas, éstas deben ser conscientes de las responsabilidades personales y legales que adquieren al ser evaluadas (AERA et al., 2014). Por ejemplo, la divulgación de material para que otros evaluados tengan información previa sobre el cuestionario que se va a emplear, a parte de una posible infracción de los derechos de autor del cuestionario, supone una grave amenaza a la validez de las inferencias que se realicen a partir del resultado obtenido en la evaluación. Por ello, aunque no sea una responsabilidad directa de los psicólogos, sí debería realizarse una tarea didáctica en la que se

muestre la importancia de un comportamiento responsable ante la evaluación. La importancia de la conducta de la persona evaluada, y su amenaza a la validez de las inferencias realizadas, puede apreciarse en el elevado número de artículos científicos que tratan de detectar las posibles conductas irregulares realizadas éstos (e.g., Steger et al., 2021; Décieux, 2022; Ranger et al., 2022; Schultz et al., 2022).

El cuestionario

De los instrumentos de evaluación empleados se espera que posean determinadas características que justifiquen su utilización. Por ejemplo, las propiedades psicométricas deben ser buenas (se debe evaluar con una precisión adecuada, el cuestionario debe haber mostrado evidencias de validez…), la baremación debe ser actual, la muestra empleada tanto para la baremación como para el cálculo de las propiedades psicométricas debe ser la adecuada teniendo en mente el uso que se pretende realizar…

Toda la información sobre la calidad del cuestionario debe aparecer en el manual, por lo que, al ser proporcionadas por la casa editorial responsable de la creación o adaptación de la prueba, corre el riesgo de tener cierto nivel de sesgo. Por ello, al igual que ocurre en países de nuestro entorno (e.g., Reino Unido, Holanda…), la Comisión de Test del Colegio Oficial de Psicólogos ha diseñado un modelo estandarizado de evaluación que permite a los usuarios tener conocimiento de la calidad técnica de los cuestionarios. En este proceso se implica tanto a las casas editoriales de los test (sin cuya ayuda este proceso no podría desarrollarse), expertos en la materia sustantiva de la prueba a evaluar, expertos en Psicometría, y, como no, a la Comisión Nacional de Test.

Para aquellos lectores interesados en las evaluaciones realizadas en otros países se les recomienda la lectura de Evers (2012) o, por ejemplo, acudir a la página web de Buros Center for Testing (http://www.buros.org) en la que también se pueden encontrar evaluaciones realizadas a pruebas en español.

En este artículo se presentan los resultados obtenidos en la novena evaluación nacional de los test editados en España. En total, tras esta evaluación, se ha valorado un total de 89 pruebas desde su comienzo en 2012. Los resultados de las evaluaciones realizadas son públicos y están a libre disposición en la página web del Colegio Oficial de Psicólogos en la siguiente dirección web: https://www.cop.es/index.php?page=evaluacion-tests-editados-en-espana. Así mismo, en la misma página, se puede acceder a cada uno de los informes generales realizados sobre cómo se llevó a cabo cada una de las evaluaciones anuales y las conclusiones generales que se pueden extraer (Elosua & Geisinger, 2016; Fonseca-Pedrero & Muñiz, 2017; Gómez-Sánchez, 2019; Hernández et al., 2015; Hidalgo & Hernández, 2019; Muñiz et al., 2011; Ponsoda & Hontangas, 2013; Viladrich et al., 2020).

Método

Participantes

Para realizar las evaluaciones de las diferentes pruebas seleccionadas se contactó con 12 profesores universitarios para que hiciesen las funciones de revisores. La selección de éstos se realizó tratando de asegurar la equidad de género, que estuviesen representadas el mayor número de universidades españolas, así como que las personas seleccionadas fuesen expertas en el constructo evaluado o en Psicometría, y que no existiese ningún tipo de conflicto de intereses. Cada prueba fue evaluada por dos revisores, uno experto en la materia sustantiva y otro en Psicometría. En la Tabla 1 se lista a los revisores que colaboraron en la realización de la evaluación.

Tabla 1.
Listado de revisores participantes en la novena evaluación de test

Nombre y Apellidos	Filiación
Juana María Breton López	Universidad Jaume I
Ramón Fernández Pulido	Universidad de Salamanca
María José Fernández Serrano	Universidad de Granada
Carmen García García	Universidad Autónoma de Madrid
Arantxa Gorostiaga Manterola	Universidad del País Vasco
Nicolás Gutiérrez Palma	Universidad de Jaén
Francisco Pablo Holgado Tello	UNED
Francisco Javier del Rio Olvera	Universidad de Cádiz
María Soledad Rodríguez González	Universidad de Santiago de Compostela
Elena Rodríguez Naveiras	Universidad Europea Canarias
Manuel Jesús Ruiz Muñoz	Universidad de Extremadura
Inmaculada Valor-Segura	Universidad de Granada

Instrumento

CET-R. El Cuestionario para la Evaluación de los Test Revisado (CET-R; Hernández et al., 2016) que se basa en el Modelo de Evaluación de Test elaborado por la European Federation of Professional Psychologists Associattions (Evers et al., 2013).

El cuestionario está formado por tres apartados diferentes precedidas de unas breves instrucciones, dirigidas a los evaluadores, en las que se informa sobre el procedimiento a seguir para cumplimentar las diferentes secciones que se exponen a continuación:

a) Descripción general del test. Está constituido por 28 ítems en los que se evalúa, tanto en un formato de respuesta cerrada como de respuesta breve las diferentes características del cuestionario evaluado (e.g., fecha de publicación, fecha de adaptación, área de aplicación, formato de los ítems, descripción de las poblaciones a las que la prueba es aplicable, precio del juego completo)
b) Valoración de las características del test. Esta sección a su vez se divide en:
   - Características generales del cuestionario (10 ítems). En ella se evalúan aspectos como la calidad de los materiales del test (objetos, material impreso o software), la fundamentación teórica, la calidad del proceso de adaptación del test, la calidad del proceso del desarrollo de los ítems, entre otros. El formato de respuesta es una escala de 5 puntos (1: Inadecuado, 2: Adecuado con carencias, 3: Adecuado, 4: Bueno, y 5: Excelente) en la que en algunos ítems se incluyen las opciones "Característica no aplicable a este instrumento" o "No se aporta información en la documentación."
   - Validez (19 ítems). En este apartado se evalúan diferentes evidencias de validez del cuestionario. El formato de respuesta es el mismo que en la sección anterior. Los ítems se reparten del siguiente modo:

- Evidencia de validez en relación con el contenido (2 ítems que evalúan tanto la calidad de la representación del contenido como el juicio de expertos realizado).
- Evidencia de validez en relación con otras variables (14 ítems). En este apartado se evalúa la relación (tanto convergente como divergente) con diferentes test, así como con un criterio externo. En esta sección, a parte de los 14 ítems mencionado anteriormente se incluyen 5 preguntas breves, en las que el evaluador debe indicar el procedimiento para la obtención de las muestras, la representatividad de las mismas …
- Evidencia de validez en relación con la estructura interna (2 ítems e los que se evalúa tanto la calidad del estudio de la estructura dimensional del cuestionario y la calidad del estudio del posible funcionamiento diferencial de los ítems).
- Acomodaciones realizadas (1 ítem). En este apartado el formato de respuesta es dicotómico (sí o no), y en caso afirmativo se debe responder una pregunta breve en la que se deben exponer cuáles han sido las acomodaciones realizadas y si se han justificado adecuadamente en el manual.

○ Fiabilidad (14 ítems). Esta subsección comienza con un ítem en el que se pregunta sobre la información aportada sobre la fiabilidad del test (tipos de coeficientes, error típico de medida, Función de Información…), para posteriormente pasar a evaluar la fiabilidad desde sus diferentes perspectivas, tanto desde la perspectiva clásica como desde la Teoría de la Respuesta a los Ítems (TRI):
- Equivalencia o formas paralelas (3 ítems)
- Consistencia Interna (3 ítems)
- Estabilidad o test-retest (2 ítems)
- Fiabilidad desde la perspectiva de la TRI (3 ítems)
- Fiabilidad interjueces (2 ítems)

○ Baremos e interpretación de las puntuaciones (9 ítems). Esta sección se divide a su vez en:
- Interpretación normativa (5 ítems)
- Test Referidos al Criterio (4 ítems)

Al finalizar cada sección (características generales, validez, fiabilidad y baremación) hay una pregunta abierta para que el evaluador plasme sus impresiones generales de manera más cualitativa. En esta sección se les solicita que indique cuales son los puntos fuertes que destacarían, así como las deficiencias que han encontrado y que se deberían solucionar.

c) Valoración global del test. En esta sección se le solicita al evaluador que en una extensión máxima de mil palabras exprese su opinión sobre los puntos fuertes y débiles del test, las recomendaciones sobre su uso en las diferentes áreas profesionales, así como las características del test que podrían ser mejoradas. Finalmente, se realiza una valoración cuantitativa de las características evaluadas calculando el promedio de las calificaciones dadas en los diferentes ítems de los distintos apartados.

El CET-R está a libre disposición en la página web del Colegio de Psicólogos (https://www.cop.es/index.php?page=evaluar-calidad).

Procedimiento

Las distintas Casas Editoriales (TEA-Hogrefe, Pearson Educación, GiuntiEOS y CEPE) junto con la Comisión Nacional de Test decidieron las diferentes pruebas que serían sometidas a evaluación. En esta novena edición se evaluaron 6 cuestionarios (ver Tabla 2).

Tabla 2.
Listado de test evaluados en la novena edición.

Acrónimo	Nombre	Editorial	Año de publicación/ revisión
DAS	Escala de Ajuste Diádico	TEA Hogrefe	2017
MacArthur	MacArthur Inventario de Desarrollo Comunicativo	TEA Hogrefe	2005
Bayley	Escala Bayley de Desarrollo Infantil III	Pearson Educación	2015
Raven's 2	Matrices Progresivas de Raven 2	Pearson Educación	2019
CAG	Cuestionario de Autoconcepto Garley	GiuntiEOS	2019
BECOLE-R	Batería de Evaluación Cognitiva de las Dificultades en Lectura y Escritura. Revisada y Renovada	CEPE	2019

Tras decidir las pruebas que serían evaluadas, la Comisión Nacional de Test encarga al coordinador de la evaluación (autor de este artículo) seleccionar a los evaluadores. Tras la aceptación de éstos se les envió tanto el CET-R en su versión electrónica como un ejemplar completo del cuestionario que debían evaluar. Los revisores aplicaron el CET-R a la prueba que se les había adjudicado y, una vez finalizado, devolvieron el CET-R cumplimentado al coordinador. La tarea de los revisores fue remunerada con 50 euros y con el cuestionario sobre el que habían realizado la evaluación. Una vez que el coordinador disponía de los dos informes realizados por los revisores los puso en común y realizó un informe preliminar para cada uno de los test. Este informe fue enviado a cada editorial responsable del cuestionario para que realizasen las alegaciones que considerasen oportunas. Tras atender a las alegaciones presentadas, el coordinador presentó los informes definitivos a la Comisión Nacional de Test.

Un resumen esquemático del procedimiento seguido puede verse en la Figura 1.

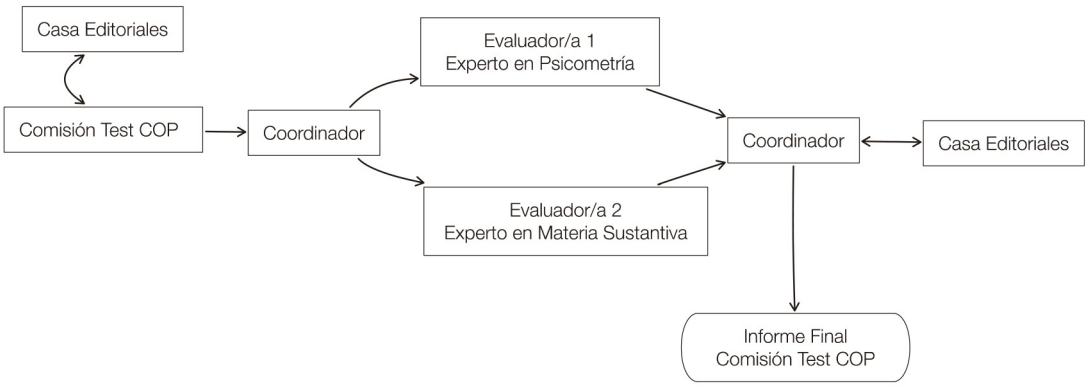
Resultados

Los informes finales individuales de las seis pruebas evaluadas en la novena edición pueden consultarse y descargarse, junto con los de las anteriores evaluaciones nacionales, en la página web del Colegio de Psicólogos (https://www.cop.es/index.php?page=evaluacion-tests-editados-en-espana).

En la Tabla 3 se muestra un resumen de las puntuaciones promedio obtenidas por cada uno de los cuestionarios en cada una de las dimensiones evaluadas. El rango de estas puntuaciones oscila entre 1 (Inadecuado) y 5 (Excelente).

En términos generales cabe destacar que la gran mayoría de puntuaciones obtenidas supera el 3,5, punto que se considera de paso entre la alternativa de respuesta Adecuado (3) y Bueno (4),

Figura 1.
Procedimiento seguido para la realización de la evaluación.



siendo la mayoría de las puntuaciones obtenidas igual o superior a 4 (Bueno). Al comparar las puntuaciones obtenidas en la novena edición con el histórico obtenido de las evaluaciones previas, se puede apreciar que los resultados obtenidos siguen la misma tendencia que los obtenidos por las pruebas evaluadas previamente. Tal y como se puede apreciar en la Tabla 3 apenas hay diferencias entre ambas puntuaciones.

En el apartado en el que se evalúa el desarrollo del cuestionario se han obtenido promedios que se pueden considerar como buenos acercándose a excelentes (superior a cuatro puntos y muy cercanos a 4,5) tanto en la calidad de los materiales y la documentación de los cuestionarios como en el proceso de adaptación cultural de los mismos. La puntuación promedio en el resto de aspectos evaluados en este componente, Fundamentación teórica y Análisis de los ítems, es buena, ya que, aunque se aproxima a los 4 puntos no llega a rebasarla.

En relación con los estudios sobre las evidencias de validez de los cuestionarios evaluados, es resaltable el hecho de que todos los cuestionarios realizan estudios de las evidencias de validez tanto de contenido como en relación con otras variables, obteniendo una buena puntuación promedio. Destaca la puntuación en las evidencias de validez en relación con el contenido, ya que es de 3,96, superando con creces el punto de corte de 3,5 necesario para ser considerada como buena.

También obtiene una puntuación que puede considerarse como adecuado el estudio que realizan del Funcionamiento Diferencial de los Ítems (DIF). Dentro de las diferentes evidencias de validez destaca el estudio que se realiza sobre la estructura interna de los cuestionarios, ya que obtiene una puntuación que se aproxima a la necesaria para ser considerada como excelente.

Respecto a la precisión en la medida se evalúan diferentes aproximaciones a la fiabilidad. Cabe destacar que ninguno de los cuestionarios evaluados en esta edición (como en ninguna otra) evalúa la fiabilidad desde la perspectiva de la equivalencia mediante formas paralelas. En el resto de los apartados, fiabilidad como consistencia interna, como estabilidad y desde el marco de la Teoría de la Respuesta a los Ítems, las puntuaciones promedio obtenidas son excelentes. El único cuestionario que evalúa la fiabilidad siguiendo un procedimiento de acuerdo interjueces es el DAS.

Finalmente, la puntuación promedio que se obtiene en la evaluación de la calidad de los baremos y las interpretaciones de las puntuaciones proporcionadas por los cuestionarios es buena. Dentro de este apartado es destacable el hecho de que el

Tabla 3.
Puntuaciones obtenidas por los test analizados en la novena evaluación

	DAS	CAG	BECOLE	BAYLEY-III	MacArthur	Raven's 2	Promedio	Histórico
Desarrollo: Materiales y documentación	4,75	3	4,5	5	4,5	5	4,46	4,3
Desarrollo: Fundamentación teórica	4	3	3,5	4,5	4	4,5	3,92	4,1
Desarrollo: Adaptación	5	--	--	3	4,5	5	4,38	4,3
Desarrollo: Análisis de los ítems	5	4	4,5	2	--	3	3,70	3,8
Validez: contenido	4	3	4	4	4,5	4,25	3,96	3,8
Validez: relación con otras variables	4,5	3	3	3	4	4,25	3,63	3,6
Validez: estructura interna	--	4,5	4,5	3,5	--	--	4,17	3,7
Validez: análisis del DIF	--	4	4	2	--	--	3,33	--
Fiabilidad: equivalencia	--	--	--	--	--	--	--	--
Fiabilidad: consistencia interna	5	3,5	4,5	5	5	4,5	4,58	4,2
Fiabilidad: estabilidad	4	--	--	3	4	5	4	3,5
Fiabilidad: TRI	--	4	4	--	--	4,5	4,17	--
Fiabilidad: interjueces	3	--	--	--	--	--	--	--
Baremos e interpretación de las puntuaciones	3,5	4	4	2,5	4	4,75	3,79	4,1

Nota. El rango de las puntuaciones oscila entre 1 y 5 siendo 1 = inadecuado; 2 = Adecuado con carencias; a partir de 2,5 = Adecuado; a partir de 3,5 = Bueno; A partir de 4,5 = Excelente. El símbolo -- indica que no se aporta información o no procede

cuestionario BAYLEY-III obtiene la puntuación más baja con 2,5 puntos (arrastrando hacia abajo la media total de las pruebas analizadas en esta evaluación). Esta puntuación se debe a que, a pesar del elevado tamaño muestral empleado por la prueba para realizar los baremos, el origen de los participantes es, en su gran mayoría, estadounidense, lo que penaliza negativamente la calidad de los baremos obtenidos.

A pesar de que en la presente evaluación tres cuestionarios evalúan tanto el DIF como la fiabilidad empleando algún procedimiento de la TRI, no se disponen de medias históricas con las que hacer una comparativa en estos apartados. Esto se debe a que, aunque en ediciones anteriores algunas pruebas evaluadas evaluaban estos aspectos, el número de datos obtenidos es aún insuficiente. Por ello, es recomendable que se haga un mayor énfasis en estos aspectos en la elaboración y estudios de validez de los tests publicados (Gómez-Sánchez, 2019; Muñiz y Fonseca-Pedrero, 2019).

Conclusiones

En términos generales puede concluirse que los resultados obtenidos en esta edición son similares a los obtenidos en las evaluaciones previas. Esto, lejos de ser una limitación o un indicador de que no se está mejorando la calidad de los cuestionarios editados en España, ahonda en la alta calidad de los test evaluados hasta el momento, ya que las puntuaciones obtenidas en las diferentes dimensiones evaluadas por el CET-R son sistemáticamente buenas o excelentes. Así mismo, también es resaltable el hecho de que cuando la prueba evaluada es una adaptación, ésta se ha realizado siguiendo las Directrices de la Comisión Internacional de Test que además de mejorar las adaptaciones de los cuestionarios también permiten una mejor comparación de las puntuaciones entre distintas culturas (Hernández et al., 2020; International Test Commission, 2018; Muñiz et al., 2013).

Aunque, tal y como se expuso anteriormente, la calificación general de los cuestionarios evaluados es buena o excelente, cabe destacar que las puntuaciones más altas se obtienen en el apartado de fiabilidad. Ésta es evaluada por la totalidad de las pruebas como consistencia interna (calculando tanto α como ω), si bien, también es evaluada como estabilidad temporal (siguiendo el procedimiento de test-retest) en cuatro de las seis pruebas, y la mitad de las pruebas evalúa la fiabilidad siguiendo algún procedimiento enmarcado dentro de la TRI. Ningún cuestionario considera la fiabilidad como equivalencia, lo cual es totalmente comprensible si atendemos a la dificultad de elaborar dos test paralelos para su cálculo.

En relación con la dimensión validez del CET-R destaca que todos los cuestionarios evaluados comprueban las evidencias de validez de contenido y en relación con otras variables. La mayor puntuación (4,17) se obtiene en las evidencias de validez en relación con la estructura interna (tanto en la versión exploratoria como confirmatoria). Como aspecto a mejorar en este apartado se podría señalar el hecho de que apenas se realizan estudios en los que se evalúe el posible funcionamiento diferencial de los ítems (solo lo han comprobado tres de los seis cuestionarios evaluados).

El CET-R es una herramienta muy útil a la hora de evaluar la calidad de las pruebas, si bien no está libre de problemas que deben afrontarse para tratar de obtener las mejores evaluaciones posibles. Uno de los problemas que se ha encontrado es el hecho de que existen ítems excesivamente rígidos para capturar aquello que pretenden medir. Por ejemplo, en los ítems en los que se evalúa el cuestionario en función de los tamaños de las correlaciones promedio entre el test y un criterio (i.e., ítem 2.11.2.2.6) se obtiene una puntuación de excelente si la correlación es igual o superior a 0,55. El valor de la correlación está directamente relacionado con la dispersión de la muestra, por lo que si, por ejemplo, la muestra es clínica y por ende muy homogénea, la correlación que se obtendrá será baja, implicando esto que en este ítem del CET-R el cuestionario obtendrá una puntuación más baja de la que realmente merecería. Una posible solución a este problema es que al igual que se emplean preguntas cualitativas en los casos de los tamaños muestrales, permitiendo al evaluador justificar que se empleen tamaños pequeños, estas preguntas se debería incluir en algunos ítems del CET-R para realizar una evaluación más justa de los cuestionarios.

Otro punto de mejora que puede tener el CET-R es el hecho de que la valoración realizada por los jueces no es coincidente. Hay que tener en cuenta que la evaluación es realizada tanto por un especialista en la materia que el test evalúa como por un especialista en psicometría. Si bien esto es un punto fuerte, ya que permite tener una visión tanto desde una perspectiva sustantiva como psicométrica del test a evaluar, en ocasiones, al igual que ocurre en todo procedimiento de evaluación por pares, lleva a la existencia de desacuerdos entre los evaluadores. Este problema no es algo exclusivo del CET-R, ya que los niveles de acuerdo entre jueces en evaluaciones de este tipo pueden considerarse sistemáticamente bajos (Hogan et al., 2021). Por ello, es de vital importancia que el responsable de la evaluación trate de solucionar estas diferencias, por ejemplo, buscando un tercer evaluador o mediante un procedimiento de reuniones de conciliación entre jueces. Sin ningún lugar a dudas, cualquier intento de incrementar los niveles de concordancia entre los evaluadores de los test ahondará en la validez de las conclusiones extraídas a partir de las valoraciones de los mismos.

La realización de las evaluaciones nacionales tiene como objetivo mejorar la calidad de los test empleados, el uso que se hace de los mismos, y con ello, la práctica profesional (Elosua & Geisinger, 2016). Sin ningún lugar a dudas, se puede considerar que la calidad de las pruebas evaluadas se puede establecer entre buena y excelente, lo que llevará a que los profesionales de la psicología puedan realizar mejores evaluaciones y con ello tomen mejores decisiones. Si la medida de lo psicológico es la base de la labor de los profesionales de la psicología, disponer de buenos instrumentos de medida es vital para que el trabajo posterior sea coherente y esté fundamentado.

Agradecimientos

Me gustaría mostrar públicamente mi más sincero agradecimiento al personal de administración del COP, a los miembros de la Comisión Nacional de Test, a los evaluadores implicados y a las casas editoriales TEA Hogrefe, Pearson Educación, GiuntiEOS y CEPE, ya que sin ellos estas evaluaciones no serían posibles.

Conflicto de Intereses

No existe conflicto de intereses

Referencias

American Educational Research Association, American Psychological Association, y National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. https://www.apa.org/science/programs/testing/standards.aspx

Décieux, J. P. (2022). Sequential on-device multitasking within online surveys: A data quality and response behavior perspective. *Sociological Methods and Research.* http://doi.org/10.1177/00491241221082593

Elosua, P., y Geisinger, K. F. (2016). Cuarta evaluación de test editados en España: Forma y fondo. *Papeles del Psicólogo/Psychologist Papers*, *37*(2), 82–88. https://www.papelesdelpsicologo.es/pdf/2693.pdf

Evers, A. (2012). The internationalization of test reviewing: Trends, differences, and results. *International Journal of Testing, 12*, 136–156. https://doi.org/10.1080/15305058.2012.658932

Evers, A., Muñiz, J., Hagemeister, C., Hstmælingen, A., Lindley, P., Sjöberg, A., y Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema*, *25*, 283–291. https://doi:10.7334/psicothema2013.97

Fonseca-Pedrero, E., y Muñiz, J. (2017). Quinta evaluación de test editados en España: mirando hacia atrás, construyendo el futuro. *Papeles del Psicólogo/Psychologist Papers*, *37*(1), 161–168. https://doi.org/10.23923/pap.psicol2017.2844

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

Gonzalez, O., MacKinnon, D. P., y Muniz, F. B. (2021). Extrinsic convergent validity evidence to prevent Jingle and Jangle fallacies. *Multivariate Behavioral Research*, *56*(1), 3–19. https://doi.org/10.1080/00273171.2019.1707061

Hernández, A., Hidalgo, M. D., Hambleton, R. K., y Gómez-Benito, J. (2020). International Test Commission guidelines for test adaptation: A criterion checklist. *Psicothema*, *3*(32), 390–398. https://doi.org/10.7334/psicothema2019.306

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G., y Elosua, P. (2016). Revisión del modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo/Psychologist Papers*, *37*, 192–197. https://www.papelesdelpsicologo.es/pdf/2775.pdf

Hernández, A., Tomás, I., Ferreres, A., y Lloret, S. (2015). Tercera evaluación de test editados en España. *Papeles del Psicólogo/ Psychologist Papers*, *36*(1), 1–8. https://www.papelesdelpsicologo.es/pdf/2484.pdf

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

Hogan, T., DeStefano, M., Gilby, C., Kosman, D., y Peri, J. (2021). Reviewing the test reviews: Quality judgments and reviewer agreements in the Mental Measurements Yearbook. *Applied Measurement in Education*, *34*(2), 75–84. https://doi.org/10.1080/08957347.2021.1890742

International Test Commission (2018). ITC Guidelines for Translating and Adapting Tests. *International Journal of Testing*, *18*, 101–134. https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

Muñiz, J., Elosua, P., y Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los test: Segunda edición. *Psicothema*, *25*, 151–157. https://doi.org/10.7334/psicothema2013.24

Muñiz, J., Fernández-Hermida, J. R., Fonseca-Pedrero, E., Campillo-Álvarez, Á., y Peña-Suárez, E. (2011). Evaluación de test editados en España. *Papeles del Psicólogo/Psychologist Papers*, *2*(32), 113–128. https://www.papelesdelpsicologo.es/pdf/1947.pdf

Muñiz, J., y Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test. *Psicothema, 31*(1), 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/Psychologist Papers*, *1*(41), 1–15. https://doi.org/10.23923/pap.psicol2020.2921

Ponsoda, V., y Hontangas, P. (2013). Segunda evaluación de tests editados en España. *Papeles del Psicólogo/Psychologist Papers*, *34*(2), 82–90. https://www.papelesdelpsicologo.es/pdf/2232.pdf

Ranger, J., Schmidt, N., y Wolgast, A. (2022). Detecting cheating in large-scale assessment: The transfer of detectors to new tests. *Educational and Psychological Measurement.* https://doi.org/10.1177/00131644221132723

Schultz, M., Lim, K. F., Goh, Y. K., y Callahan, D. L. (2022). OK google: What's the answer? characteristics of students who searched the internet during an online chemistry examination. *Assessment and Evaluation in Higher Education, 47*(8), 1458-1474. https://doi.org/10.1080/02602938.2022.2048356

Steger, D., Schroeders, U., y Wilhelm, O. (2021). Caught in the act: Predicting cheating in unproctored knowledge assessment. *Assessment, 28*(3), 1004-1017. https://doi.org/10.1177/1073191120914970

Viladrich, C., Doval, E., Penelo, E., Aliaga, J., Espelt, A., García-Rueda, R., y Angulo-Brunet, A. (2020). Octava evaluación de test editados en España: Una experiencia participativa. *Papeles del Psicólogo/ Psychologist Papers*, *42*(1), 1–9. https://doi.org/10.23923/pap.psicol2020.2937