

Artículo

Duodécima Evaluación de Test Editados en España

Patricia Recio-Saboya 

Universidad Nacional de Educación a Distancia (UNED), España

INFORMACIÓN

Recibido: Octubre 24, 2025

Aceptado: Febrero 9, 2026

Palabras clave

Evaluación de test

Propiedades psicométricas
CET-R

RESUMEN

Desde 2010, la Comisión de test del Consejo General de la Psicología realiza anualmente una evaluación de los test publicados en España, con el propósito de ofrecer información rigurosa y accesible acerca de su calidad. El objetivo de este trabajo es presentar los resultados de la Duodécima Evaluación de Test Editados en España. En esta edición se evaluaron seis test de cuatro editoriales y un test no comercial, diseñados para medir variables como motivación, inteligencia general, deterioro cognitivo, bienestar emocional, personalidad en adolescentes, conciencia fonológica e inteligencia emocional en niños. Como en ediciones anteriores, se siguió un proceso sistemático de revisión por pares, utilizando el *Cuestionario para la Evaluación de Test Revisado (CET-R v1.1)*, se integraron las evaluaciones en un informe provisional y, teniendo en cuenta las alegaciones de las editoriales, se redactó el informe final. Los resultados siguen la misma línea de evaluaciones anteriores, mostrando que, en términos generales, la calidad de los test es adecuada, destacando la calidad de los materiales y documentación, las evidencias de validez de contenido y la consistencia interna, así como la calidad de los baremos. De cara al futuro sería deseable mejorar la consistencia entre revisores, y fomentar la formación continua en psicometría.

Twelfth Review of Tests Published in Spain

ABSTRACT

Since 2010, the Test Commission of the General Council of Psychology has conducted an annual evaluation of the quality of tests published in Spain, with the aim of providing rigorous and accessible information regarding their quality. The purpose of this study is to present the results of the Twelfth Evaluation of Test Published in Spain. In this edition, six test from four different publishers and one non-commercial test were assessed. These instruments were designed to measure variables such as motivation, general intelligence, cognitive impairment, emotional well-being, adolescent personality, phonological awareness, and children's emotional intelligence. As in previous editions, a systematic peer review process was followed, employing the *Test Evaluation Questionnaire - Revised (CET-R v1.1)*. The evaluations were integrated into a provisional report and, after considering the claims of the publishers, a final report was prepared. The results are consistent with those of previous rounds, showing that, overall, the quality of the evaluated test is adequate, with particular strengths in the quality of materials and documentation, the evidence of content validity and internal consistency, and the quality of the norms. Looking ahead, it would be desirable to improve consistency among reviewers and to promote ongoing training in psychometrics.

Keywords

Test evaluation

Psychometric properties
CET-R

En los distintos ámbitos de la Psicología, la evaluación mediante test constituye una práctica habitual, empleándose las puntuaciones derivadas de estos instrumentos para fines tan relevantes como el diagnóstico, la toma de decisiones, la orientación, la intervención o la selección de personal, entre otros.

La calidad psicométrica de los test es un aspecto crucial para garantizar que las decisiones basadas en sus resultados sean válidas, fiables y éticas. Un test que carezca de evidencias suficientes de validez, fiabilidad y adecuación normativa o relevancia cultural puede conducir a interpretaciones erróneas y, en consecuencia, a decisiones con efectos negativos para las personas evaluadas. Además, es importante tener en cuenta que ningún instrumento de medición psicológica es “universalmente válido” para cualquier uso, en cualquier momento y con cualquier población, ya que la validez depende del contexto, el propósito y las características de la muestra (Cohen y Swerdlik, 2018).

Aunque el mayor peso en el análisis de la calidad psicométrica de un test recaiga en el constructor de la prueba, en la séptima edición de los *Standards for Educational and Psychological Testing* (AERA, APA y NCME, 2014), marco de referencia en cuanto a criterios para el desarrollo y la evaluación de pruebas, se destaca claramente la responsabilidad del profesional a la hora de seleccionar un test que mida de manera adecuada la variable de interés y cuente con una calidad métrica acreditada. El estándar 9.0 establece que el usuario del test debe conocer la evidencia de validez que sustenta las interpretaciones de las puntuaciones y asumir las consecuencias de su uso. Además, el estándar 9.1 especifica que solo individuos con la formación, credenciales y experiencia necesarias deben interpretar y aplicar los test, garantizando así su uso apropiado y ético. Por tanto, los estándares hacen explícito que la responsabilidad de un uso adecuado y ético no recae únicamente en los constructores o editores de los test, sino también en los psicólogos que los seleccionan y aplican.

En este contexto, disponer de información rigurosa y actualizada sobre la calidad de las pruebas es clave para que los profesionales de la Psicología puedan tomar decisiones informadas acerca de su utilización, ya que tendrán que analizar de manera crítica la calidad psicométrica de los test antes de su selección y aplicación. Para cubrir esta necesidad, en España se ha desarrollado un modelo sistemático de evaluación de test, el *Cuestionario para la Evaluación de Test* (CET; Prieto y Muñiz, 2000) y su versión revisada (CET-R; Hernández et al., 2016), que están basados en el modelo de revisión de test de la Federación Europea de Asociaciones de Psicología (EFPA; Evers et al., 2013).

Además de las evaluaciones técnicas, resulta fundamental conocer la opinión de los propios profesionales de la Psicología sobre el uso de los test. Con este propósito se han realizado tres encuestas nacionales a psicólogos colegiados. Los resultados de la tercera encuesta (Muñiz et al., 2020), en la que participaron 1.248 profesionales, evidencian que los test se emplean de forma habitual en la práctica psicológica y que los psicólogos no cuestionan su utilidad cuando son aplicados de manera adecuada. No obstante, los resultados también muestran una clara demanda de mayor control y regulación en su uso, así como la necesidad de una formación más sólida en psicometría y evaluación.

Es en este marco donde cobra especial importancia contar con evaluaciones independientes y actualizadas de la calidad técnica de los test. En España, dicho propósito ha sido impulsado por la

Comisión de Test del Consejo General de la Psicología (COP). Desde 2010, esta Comisión ha venido desarrollando un proyecto sistemático de evaluación de los test editados en España, habiendo ya once evaluaciones previas.

El objetivo principal de este trabajo es presentar los resultados obtenidos en la duodécima evaluación nacional de los test editados en España. La finalidad última es ayudar a los profesionales en la toma de decisiones de uso de los test, ya que esta iniciativa busca ofrecer a los profesionales información detallada sobre la calidad de los instrumentos que orienten su selección y uso adecuado (Prieto y Muñiz, 2000; Muñiz et al., 2011).

En la actualidad se supera el centenar de test evaluados, cuyos informes están disponibles para consulta y descarga gratuita en la página web del Consejo General de la Psicología en España (<https://www.cop.es/test/>). En este mismo portal se incluyen los enlaces a los artículos que documentan el proceso de evaluación de cada edición, publicados en *Papeles del Psicólogo* desde la primera edición en 2010, lo que permite seguir la evolución histórica de este proyecto a través de las diferentes coordinaciones (Muñiz et al., 2011; Ponsoda y Hontangas, 2013; Hernández et al., 2015; Elosua y Geisinger, 2016; Fonseca-Pedrero y Muñiz, 2017; Hidalgo y Hernández, 2019; Gómez, 2019; Viladrich et al., 2021; Lozano, 2023; Abad, 2024; Guilera y Barrios, 2025).

Método

Test Evaluados

En esta duodécima edición se han evaluado siete test de cuatro editoriales de referencia, TEA, Pearson, Habilmind y Giunti Psychometrics, además de un instrumento no comercial, el Cuestionario de Pfeiffer de evaluación de deterioro cognitivo (SPMSQ). Estos test abordan un amplio rango de variables psicológicas, entre las que se incluyen la motivación, la inteligencia general, los rasgos de personalidad en adolescentes, el bienestar emocional, el deterioro cognitivo, la conciencia fonológica y la inteligencia emocional (ver Tabla 1). Además, presentan aplicaciones en diversos ámbitos profesionales, como la psicología educativa, clínica, forense, del trabajo y de las organizaciones, así como en investigación.

El cuestionario Análisis del Perfil Motivacional (APM) (Valderrama et al., 2015) es un instrumento de autoinforme compuesto por 80 ítems con formato Likert, organizados en diez escalas que conforman la denominada rueda de motivos (Valderrama, 2010; 2018). Estas escalas se agrupan en cinco pares complementarios: afiliación-autonomía, cooperación-poder, hedonismo-logro, seguridad-exploración y conservación-contribución. Se orienta a la evaluación de los motivos que influyen en el rendimiento y en la conducta laboral, con aplicaciones en contextos de orientación profesional, crecimiento personal, selección de personal y gestión del talento y la diversidad.

La adaptación española del Inventario Clínico para Adolescentes Millon-II (MACI-II) (Hernández et al., 2023) es un instrumento autoinformado dirigido a la evaluación de rasgos de personalidad y síndromes clínicos en adolescentes de 13 a 18 años. La prueba integra tres tipos de información: rasgos de personalidad (patrones estables), preocupaciones subjetivas (malestar psicológico) y síndromes clínicos (de carácter agudo o transitorio. Consta de 160 ítems de respuesta verdadero/falso que requieren aproximadamente

Tabla 1

Listado de los Instrumentos de Medida Analizados en la Duodécima Evaluación de Test

Acrónimo	Nombre	Editorial	Variable de interés	Área de aplicación	Soporte
APM	Análisis de Perfiles Motivacionales	TEA	Motivación	Psicología educativa, Psicología del trabajo y las organizaciones	Papel y lápiz e informatizado
Matrices-TAI	Test Adaptativo de Inteligencia General (online)	TEA	Inteligencia general	Psicología clínica, Psicología forense, Psicología del trabajo y las organizaciones	Informatizado
MACI-II	Inventario Clínico para adolescentes Millon-II	Pearson	Rasgos de personalidad y síndromes clínicos en adolescentes	Psicología clínica, Psicología educativa, Psicología forense, investigación	Papel y lápiz e informatizado
SBE	Screening de bienestar emocional para niños y adolescentes	Habilmind	Bienestar emocional	Psicología educativa	Informatizado
SPMSQ	Cuestionario de Pfeiffer de evaluación de deterioro cognitivo	No comercial	Deterioro cognitivo	Neuropsicología, escalas clínicas	Administración oral
TECO	Prueba para la evaluación del conocimiento fonológico	Giunti Psychometrics	Conciencia fonológica	Psicología educativa	Papel y lápiz / administración oral
THINEME	Test de Inteligencia Emocional	Giunti Psychometrics	Inteligencia emocional	Psicología educativa Psicología clínica	Papel y lápiz e informatizado

25 minutos de aplicación. Sus usos incluyen la planificación del tratamiento psicoterapéutico, la evaluación neuropsicológica y psicoeducativa, así como su utilización en investigación.

El Test Adaptativo de Inteligencia General (Matrices-TAI) (Abad et al., 2020) es una versión informatizada y adaptativa del test de Matrices de TEA (Sánchez-Sánchez, Santamaría y Abad, 2015), diseñado para evaluar la inteligencia general mediante razonamiento abstracto no verbal. Puede aplicarse en una amplia franja edad (6 a 74 años) y en contextos clínicos, forenses, organizacionales y educativos. Su formato, basado en ítems de matrices gráficas incompletas (3x3), minimiza la influencia del lenguaje y está disponible en español, catalán e inglés, incorporando ayudas visuales y auditivas.

El Screening de Bienestar Emocional (SBE) (Blanco et al., 2023) es un cuestionario informatizado de 65 ítems tipo Likert, diseñado para evaluar el bienestar emocional en población infantil y adolescente (3-18 años). Evalúa trece subdimensiones agrupadas en tres bloques: emociones causales (gestión de miedo, tristeza, ira y ansiedad/estrés), factores de protección emocional (optimismo, autoestima, motivación y expresión emocional) y manifestaciones conductuales (alteraciones del sueño, abuso de pantallas, agresividad/disruptividad y aislamiento social). Ofrece versiones adaptadas por nivel evolutivo: de 3 a 8 años cumplimentado por padres o cuidadores, y de 9 a 18 años en formato de autoinforme.

El Cuestionario de Pfeiffer para detectar la existencia de deterioro cognitivo (SPMSQ-VE) (Martínez de la Iglesia et al., 2001) es la adaptación española del Short Portable Mental Status Questionnaire (SPMSQ) (Pfeiffer, 1975). Es un instrumento breve de cribado del deterioro cognitivo en personas mayores de 65 años. El cuestionario consta de 10 ítems de respuesta abierta que evalúan orientación, memoria, información sobre hechos cotidianos y capacidad de cálculo seriado. La puntuación se basa en el número de errores cometidos, con un rango de 0 a 10. Es de aplicación sencilla y rápida por profesionales sanitarios en atención primaria, y ha sido recomendado por el Sistema Nacional de Salud como prueba de referencia en la detección de demencias.

El Test de Evaluación de la Conciencia Fonológica (TECO) (Ramos et al., 2023) es una revisión del PECO (Ramos y Cuadrado, 2006) orientado a la evaluación de la conciencia fonológica en el

proceso inicial de aprendizaje de la lectura en español. Evalúa los niveles silábico y fonémico mediante seis actividades (22 ítems) que implican tareas de identificación, adición y omisión de unidades fonológicas. Las dos primeras actividades pueden aplicarse de forma colectiva y las cuatro restantes se aplican de forma individual. Su población objetivo es el alumnado del último curso de Educación Infantil, primer ciclo de Educación Primaria y en cursos superiores cuando existen dificultades de lectoescritura. Su uso principal es educativo, aunque también tiene aplicaciones en investigación psicolingüística y psicoeducativa.

El Test de Inteligencia Emocional (THINEME) (Merchán et al., 2023) es una prueba de ejecución destinada a evaluar la inteligencia emocional en población infantil (8-12 años), siguiendo el modelo de Mayer y Salovey (1997). El test se compone de 30 ítems y proporciona una puntuación global y puntuaciones específicas en cuatro ramas: percepción, facilitación, comprensión y regulación emocional. Las posibilidades de utilización van desde la orientación educativa, al ámbito clínico, psicopedagógico, y detección de perfiles emocionales en niños y niñas.

Revisores Participantes

En esta edición se contactó con 18 posibles revisores, de los cuales cuatro rechazaron participar en el proceso por distintos motivos, por lo que finalmente participaron 14 revisores, dos por cada prueba evaluada (ver Tabla 2). La selección de los revisores se realizó teniendo en cuenta que cada test debía ser evaluado por dos expertos, uno en Psicometría y otro en el área específica abordada por el instrumento. Cuando fue posible, se procuró que el experto en Psicometría contara también con publicaciones relacionadas con la variable medida por el test, y que el especialista en el área evaluada por el test hubiera publicado trabajos en los que se evalúen las propiedades psicométricas de alguna prueba.

Todos los revisores pertenecen al ámbito universitario. La mayoría adscritos al área de Metodología de las Ciencias del Comportamiento, aunque, debido al ámbito de aplicación de los test, también participaron evaluadores del área de Personalidad, Evaluación y Tratamientos Psicológicos, de Psicología Básica y de Psicología Evolutiva y de la Educación.

Tabla 2

Listado de Revisores de la Duodécima Evaluación de Test

Revisor/a	Departamento
Jesús M ^a Alvarado Izquierdo	Psicobiología y Metodología en Ciencias del Comportamiento de la Universidad Complutense de Madrid (UCM)
Isabel Benítez Baena	Metodología de las Ciencias del Comportamiento de la Universidad de Granada (UGR)
Sergio Escorial Martín	Psicobiología y Metodología en Ciencias del Comportamiento de la Universidad Complutense de Madrid (UCM)
Silvia Fernández Rivas	Psicología Básica, Psicobiología y Metodología de las Ciencias del Comportamiento de la Universidad de Salamanca (USAL)
Juana Gómez Benito	Psicología Social y Psicología Cuantitativa de la Universidad de Barcelona (UB)
Isabel Gómez Veiga	Psicología Evolutiva y de la Educación de la Universidad Nacional de Educación a Distancia (UNED)
Marcela Paz González Brignardello	Psicología de la Personalidad, Evaluación y Tratamiento Psicológico de la Universidad Nacional de Educación a Distancia (UNED)
Georgina Guilera Ferré	Psicología Social y Psicología Cuantitativa de la Universidad de Barcelona (UB)
Elena Navarro González	Psicología de la Personalidad, Evaluación y Tratamiento Psicológico de la Universidad de Granada (UGR)
Francisco José Rivera de los Santos	Psicología Experimental de la Universidad de Sevilla (US)
Encarnación Sarriá Sánchez	Metodología de las Ciencias del Comportamiento de la Universidad Nacional de Educación a Distancia (UNED)
Miguel Ángel Sorrel Luján	Psicología Social y Metodología de la Universidad Autónoma de Madrid (UAM)
Purificación Sierra García	Psicología Evolutiva y de la Educación de la Universidad Nacional de Educación a Distancia (UNED)
Rodrigo Schames Kreitchmann	Metodología de las Ciencias del Comportamiento de la Universidad Nacional de Educación a Distancia (UNED)

Descripción del Instrumento Utilizado Para la Evaluación

Para realizar las evaluaciones de los test se utilizó 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 Psychologists' Associations* (EFPA, [Evers et al., 2013](#)).

El cuestionario consta de tres bloques principales:

1. Descripción general del test: compuesto por 28 ítems cuyo objetivo es obtener información básica y objetiva sobre el instrumento, incluyendo datos editoriales, autores, fechas de publicación, área de aplicación, formato de ítems, poblaciones destinatarias, modalidades de aplicación y corrección, escalas empleadas, documentación disponible y precios de los materiales.
2. Valoración de las características del test: examina en detalle la fundamentación y propiedades psicométricas de la prueba. Este bloque se organiza en cuatro apartados, finalizando cada uno de ellos con una pregunta abierta para sintetizar las evidencias encontradas, justificar las puntuaciones, y señalar las fortalezas y debilidades de la prueba.
 - *Características generales*: analiza aspectos como calidad de materiales, documentación, fundamentación teórica, adaptación, redacción y desarrollo de ítems, así como la claridad de instrucciones y bibliografía utilizada.
 - *Validez*: se evalúan las evidencias disponibles basadas en el contenido, en la relación entre las puntuaciones del test y otras variables (incluidos estudios con criterios externos, análisis de diferencias entre grupos, validez convergente, discriminante y basadas en la estructura interna del test (p. ej., análisis factorial y funcionamiento diferencial de los ítems).
 - *Fiabilidad*: incluye distintos enfoques para valorar la precisión de las puntuaciones, como equivalencia entre formas paralelas, consistencia interna, estabilidad temporal, estimaciones desde la Teoría de Respuesta a los Ítems y fiabilidad interjueces.
 - *Baremos e interpretación de puntuaciones*: evalúa la calidad y actualización de los baremos en interpretaciones normativas y la adecuación de los procedimientos empleados para establecer puntos de corte en pruebas referidas a criterio.
3. Valoración global del test: en este apartado se solicita una evaluación cualitativa, en la que se informe de las principales fortalezas, limitaciones, recomendaciones de uso y aspectos susceptibles de mejora. Asimismo, se completa una tabla resumen con los promedios obtenidos en los distintos apartados, lo que permite una valoración cuantitativa global del instrumento.

El CET-R V1.1 está disponible en la web del Colegio Oficial de Psicólogos (<https://www.cop.es/test/>).

Proceso de Evaluación

El procedimiento de evaluación seguido en esta edición fue similar al de convocatorias anteriores. La Comisión de Test, en colaboración con las editoriales participantes (Giunti Psychometrics, Habilmind, Pearson y TEA) seleccionó las pruebas que serían sometidas a evaluación. En esta duodécima edición se evaluaron seis cuestionarios de carácter comercial y, además, la Comisión de Test incluyó la evaluación de un test no comercial, el Cuestionario de Pfeiffer de Evaluación de Deterioro Cognitivo (SPMSQ).

Una vez seleccionadas las pruebas a evaluar, la Comisión de Test designó a la persona responsable de Coordinar la Evaluación de esta edición. Entre julio y septiembre de 2023 el COP envió los test provenientes de las editoriales a la coordinadora. Según fueron llegando los test se abordó su lectura para comenzar la selección de los revisores. Para cada test se buscó la colaboración de 2 expertos, uno en Psicometría y otro en el constructo o área específica evaluada por la prueba.

En octubre, la coordinadora contactó con los posibles revisores mediante un primer mensaje de correo electrónico informando sobre la prueba asignada y sobre el instrumento de revisión que debían utilizar (CET-R v1.1). Aquellos que aceptaron la colaboración recibieron posteriormente las instrucciones detalladas de la tarea (básicamente revisar la documentación del test y cumplimentar el CET-R) y el plazo de entrega, fijado para enero de 2024. El COP envió a cada revisor todo el material relacionado con la evaluación del test asignado (manual, hojas de respuesta, láminas, cuadernillos, entre otros) por correo postal.

En el caso de la prueba no comercial, la coordinadora recopiló y envió por email a los revisores los artículos relevantes sobre su desarrollo, adaptación y/o propiedades psicométricas del Cuestionario de Pfeiffer de evaluación de deterioro cognitivo para que pudieran llevar a cabo la revisión. Se consideró como artículo principal para hacer la revisión el estudio de [Martínez et al. \(2001\)](#), al ser la adaptación al español utilizada en nuestro país y tratarse de un trabajo muy completo en cuanto al estudio de sus propiedades psicométricas. Además, se envió el artículo original de [Pfeiffer \(1975\)](#), y cuatro artículos cuyo objetivo fue proporcionar información sobre la calidad psicométrica del test con muestra española, con una pequeña explicación de por qué podría ser relevante para realizar la revisión. Como material complementario se envió un listado de 37 artículos de aplicaciones del Cuestionario en muestra española, en el que estaban marcados los que aportaban algún dato relacionado con la calidad métrica de la prueba. De este último listado no se enviaron los trabajos completos porque, a priori, no aportan información adicional, pero la coordinadora quedó a disposición para enviar y comentar cualquiera de ellos.

Durante el periodo de revisión la coordinadora estuvo disponible para resolver las dudas surgidas en el proceso de revisión de los test.

Para realizar el informe provisional de cada uno de los test se integraron las evaluaciones realizadas por los dos revisores. En los apartados cualitativos del CET-R se sintetizó la información considerando las aportaciones de los dos expertos y la evaluación realizada por la coordinadora. En los apartados cuantitativos cuando existían discrepancias leves en las puntuaciones numéricas, se utilizó la media aritmética. En los casos en que las diferencias eran mayores, la coordinadora estableció la puntuación final valorando los argumentos de los revisores y la información disponible en el manual. Si se detectaba algún error evidente en una de las evaluaciones, la puntuación definitiva se obtenía mediante la media entre la calificación del otro revisor y la de la propia coordinadora.

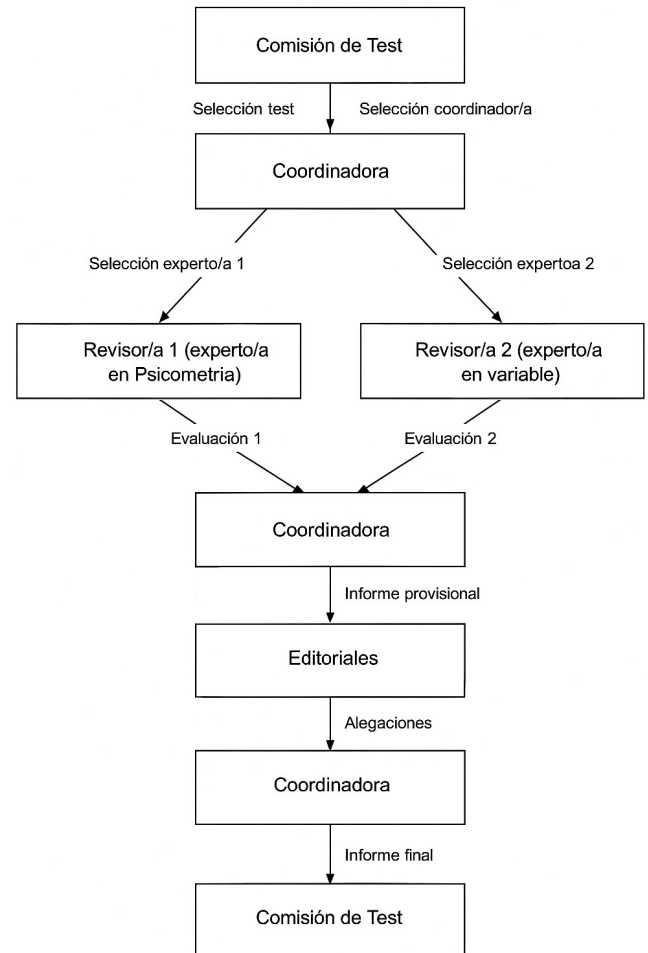
Los informes provisionales fueron enviados al COP entre marzo y abril, desde donde se remitieron a las distintas editoriales y a los autores de la adaptación (en el caso del test no comercial).

Las editoriales tuvieron un mes de plazo para realizar las alegaciones que estimaron oportunas a los informes provisionales. En algunos casos aportaron información adicional que se tuvo en cuenta para el informe final. La coordinadora revisó las alegaciones para la elaboración del informe definitivo de cada una de las pruebas. Finalmente, un miembro de la Comisión de test del COP revisó los informes finales, antes del envío definitivo para su publicación.

En la [Figura 1](#) se esquematiza el proceso de evaluación.

Figura 1

Proceso de la Duodécima Evaluación de Test Editados en España



Resultados

En la [Tabla 3](#) se recogen las valoraciones promedio obtenidas por los distintos test en cada una de las dimensiones analizadas. El rango de puntuaciones oscila entre 1 (Inadecuado) y 5 (Excelente), contemplándose además dos opciones diferenciadas para los casos en los que no se aporta información: N/A (no se aporta información, pero no es fundamental dada la finalidad del test) y 0 (no se aporta información y ésta se considera fundamental dada la finalidad del test).

En conjunto, la mayor parte de las calificaciones se sitúan por encima de 3,5, lo que las ubica en un nivel entre adecuado y bueno, siendo frecuentes las puntuaciones iguales o superiores a 4. Al comparar los resultados de esta duodécima edición con el histórico de evaluaciones previas, se observa una continuidad en la tendencia general, con ligeras variaciones en algunos apartados concretos (ver [Tabla 3](#)).

El primer bloque corresponde a los aspectos generales relacionados con el desarrollo de las pruebas. En el apartado de Materiales y documentación las valoraciones son elevadas, con una media de 4,0 y varias pruebas con puntuaciones entre buenas y excelentes (Matrices-TAI = 5,0; TECO = 4,5). En Fundamentación

Tabla 3*Puntuaciones Obtenidas por los Test Analizados en la Duodécima Evaluación*

Característica evaluada	APM	Matrices-TAI	MACI-II	SBE	SPMSQ	TECO	THINEME	Media	Histórico
Materiales y documentación	4,2	5	4,5	4	3	4,5	3	4,0	4,3
Fundamentación teórica	4	5	4	2,5	2,5	4	4	3,7	4,1
Adaptación	---	---	4,5	---	4	---	---	4,3	4,2
Análisis de ítems	3	5	---	3,5	3	4	3	3,6	3,8
Validez: Contenido	4	5	3,5	3,5	---	2,5	4	3,8	3,7
Validez: Relación con otras variables	3,1	4,5	3,5	0	3,8	3,5	2,8	3,0	3,6
Validez: Estructura interna	3,8	4,8	---	4	---	2	4	3,7	3,7
Validez: Análisis del DIF	3,5	4,5	---	3	---	---	---	3,7	4,1
Fiabilidad: Equivalencia	---	---	---	---	---	---	---	---	---
Fiabilidad: Consistencia interna	4	5	3,5	4,5	3	4	2,8	3,8	4,2
Fiabilidad: Estabilidad	---	4	3	3,5	---	---	---	3,5	3,4
Fiabilidad: TRI	---	5	---	---	---	---	---	5,0	---
Fiabilidad Interjueces	---	---	---	---	5	---	2,5	3,8	---
Baremos e interpretación puntuaciones	4	5	4	4	3,7	3,7	3	3,9	4

Nota. Las puntuaciones de la tabla siguen una escala de 0 a 5. El símbolo --- indica que no se aporta información o no procede; Histórico es la puntuación media en las ediciones anteriores.

teórica los resultados son algo más dispares, con puntuaciones que oscilan entre 2,5 y 5,0, lo que se traduce en una media de 3,7, ligeramente inferior al valor histórico. En cuanto a la Adaptación, las pruebas que ofrecen información alcanzan valores de bueno a excelente (MACI-II = 4,5; SPMSQ = 4,0).

En relación con las evidencias de validez, las puntuaciones en Validez de contenido son en su mayoría adecuadas o buenas (media = 3,8), aunque se aprecia cierta heterogeneidad entre pruebas. En la evidencia sobre la Relación con otras variables las puntuaciones están entre adecuado y bueno excepto dos pruebas (una con alta valoración y otra muy baja). Respecto a la Estructura interna, la media se sitúa en 3,7, en línea con ediciones anteriores, aunque con valores desiguales (p.ej., TECO = 2,0 frente a Matrices-TAI = 4,8). En cuanto al análisis del DIF, las valoraciones son positivas (media = 3,7), aunque no siempre se recoge información de este tipo.

En el bloque de fiabilidad, los resultados muestran que la Consistencia interna es el indicador más utilizado y mejor valorado, con una media de 3,8, aunque algunas pruebas obtienen puntuaciones solo adecuadas. Los datos de Estabilidad son más limitados y presentan valores intermedios (media = 3,5). En el caso en el que se aplican análisis con la TRI (Matrices-TAI) la calificación es excelente (5,0). También se han recogido algunas valoraciones en Fiabilidad interjueces, con resultados muy dispares.

Finalmente, en el apartado de Baremos e interpretación de puntuaciones las medias alcanzan un nivel bueno (3,9), no habiendo ningún test evaluado por debajo de la puntuación 3 (adecuado).

En conjunto, los resultados reflejan un patrón relativamente estable respecto a evaluaciones previas, como se puede apreciar en la [Tabla 4](#) y también en la [Figura 2](#), donde se presenta la evolución de los principales bloques evaluados (aspectos generales, validez, fiabilidad y baremos). Cabe destacar que hay que interpretar este

tipo de datos con cautela, ya que el número de pruebas a evaluar en cada edición no es demasiado alto, por lo que las valoraciones medias de las evaluaciones en las que ha habido alguna prueba con puntuaciones excepcionalmente altas o bajas pueden distorsionar la media de esa evaluación. Además, debe tenerse en cuenta que en las cuatro primeras evaluaciones se utilizó el CET (en lugar de la versión revisada), en la que variaban algunos epígrafes, por ejemplo, no había apartado de fiabilidad bajo la Teoría de Respuesta al Ítem.

Como se puede ver en la [Tabla 4](#), en la mitad de las evaluaciones no hay ningún test que aporte evidencias de fiabilidad utilizando formas paralelas, fiabilidad interjueces o proporcionando la función de información bajo la Teoría de Respuesta al ítem (TRI).

En la [figura 2](#) se observa que los promedios obtenidos en los cuatro bloques principales del CET-R, a lo largo de las 12 evaluaciones realizadas hasta la fecha, se sitúan en un rango de 3 (aceptable) a 4,5 (bueno-excelente) lo que refleja, en términos generales, la buena calidad de los test evaluados.

Discusión

La evaluación de la calidad de los test psicológicos es una condición imprescindible para asegurar que las decisiones que se derivan de su aplicación sean válidas, fiables y éticas. Los criterios métricos de calidad fundamentales (validez y fiabilidad) no son atributos universales ni permanentes de un instrumento, sino que dependen de las características de la población, el contexto de uso y los fines para los que se aplican ([AERA, APA y NCME, 2014](#)). Bajo esta premisa, las evaluaciones periódicas de los test publicados en España constituyen un recurso fundamental, pues permiten identificar fortalezas y limitaciones técnicas de los instrumentos disponibles, además de prevenir interpretaciones erróneas y a

Tabla 4*Puntuaciones Promedio Obtenidas por los Test Analizados en Cada una de las Evaluaciones*

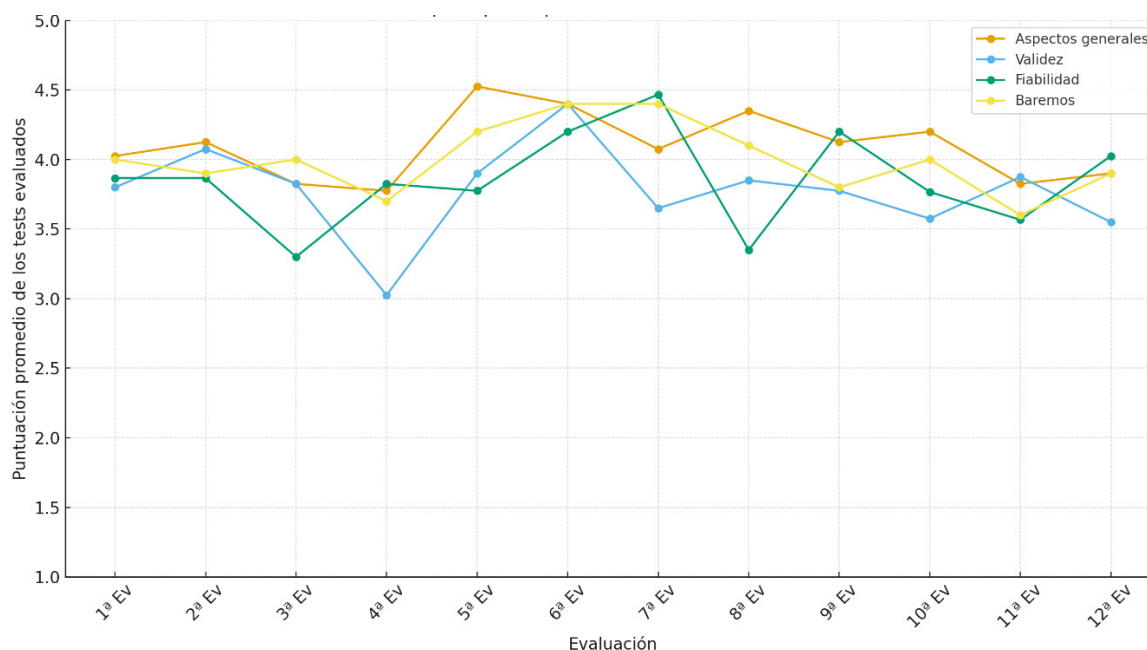
Característica evaluada	1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	10ª	11ª	12ª
Materiales y documentación	4,3	4,3	4,1	4,0	4,5	4,6	4,1	4,3	4,5	4,5	4,5	4,0
Fundamentación teórica	4,1	4,0	3,6	3,6	4,7	4,6	3,9	4,7	3,9	4,2	3,7	3,7
Adaptación	4,1	4,3	3,9	3,8	4,7	4,2	4,5	4,8	4,4	4,3	3,3	4,3
Análisis de ítems	3,6	3,9	3,7	3,7	4,2	4,2	3,8	3,6	3,7	3,8	3,8	3,6
Validez: Contenido	4,1	3,5	3,6	3,3	4,2	4,3	3,3	4,0	4,0	3,8	3,3	3,8
Validez: Relación con otras variables	3,6	3,6	3,6	3,2	3,6	3,8	3,7	3,3	3,6	3,7	3,8	3,0
Validez: Estructura interna	3,7	4,2	3,8	3,6	3,8	4,5	3,6	3,1	4,2	3,1	3,4	3,7
Validez: Análisis del DIF	---	5,0	4,3	2,0	4,0	5,0	4,0	5,0	3,3	3,7	5,0	3,7
Fiabilidad: Equivalencia	4,0	4,0	---	3,0	3,0	---	4,5	3,0	---	---	---	---
Fiabilidad: Consistencia interna	3,8	4,2	3,6	4,2	4,4	4,4	4,8	4,3	4,6	4,0	4,0	3,8
Fiabilidad: Estabilidad	3,8	3,4	3,0	3,1	3,7	3,6	4,1	3,1	4,0	3,3	3,2	3,5
Fiabilidad: TRI	---	---	---	---	---	4,8	---	---	4,2	4,0	3,5	5,0
Fiabilidad: Interjueces	---	---	---	---	5,0	4,0	---	---	3,0	4,0	---	3,8
Baremos e interpretación puntuaciones	4,0	3,9	4,0	3,7	4,2	4,4	4,4	4,1	3,8	4,0	3,6	3,9

proteger de ellas a los usuarios, lo que refuerza la confianza en los test por parte de los profesionales de la Psicología (Muñiz et al., 2020).

En la actualidad, ya se ha superado el centenar de test evaluados, una cifra que refleja la consolidación de este proyecto desde sus primeras ediciones (Muñiz et al., 2011; Ponsoda y Hontangas, 2013; Hernández, Tomás, Ferreres, y Lloret, 2015), lo que pone de manifiesto tanto el compromiso de la Comisión de Test del Consejo General de la Psicología como el interés de la comunidad científica por disponer de un sistema independiente y riguroso de revisión. Desde una perspectiva aplicada, el que los informes estén disponibles en abierto en la web del COP (<https://www.cop.es/test/>)

constituye un recurso de gran valor, al facilitar que cualquier profesional pueda acceder a información contrastada sobre la calidad psicométrica de los instrumentos antes de decidir su aplicación. Ahora bien, la consolidación del proyecto no garantiza por sí misma que los psicólogos aprovechen plenamente este recurso.

En este sentido, la encuesta realizada a profesionales de la Psicología (Muñiz et al., 2020), puso de manifiesto que, aunque los test se emplean de manera generalizada y se reconoce su utilidad, existe una clara demanda de mayor regulación en su uso y de una formación más sólida en psicometría. A la vez, se evidenció un conocimiento limitado del trabajo de la Comisión de Test, lo que

Figura 2*Puntuaciones Obtenidas en cada una de las Evaluaciones en los Cuatro Bloques Generales*

subraya la necesidad de intensificar las labores de difusión y transferencia de este recurso al ámbito profesional.

Desde el punto de vista metodológico, la utilización del Cuestionario para la Evaluación de Test (CET; Prieto y Muñiz, 2000) y de su versión revisada (CET-R; Muñiz, Hernández y Ponsoda, 2015), en todas las evaluaciones realizadas hasta el momento permite establecer comparaciones entre ediciones y detectar tendencias en la calidad de los test editados en España. No obstante, los revisores han señalado ciertas dificultades prácticas en la aplicación del instrumento. Una de las más destacadas se refiere a la valoración de los tamaños muestrales en contextos complejos, especialmente cuando la utilización de varias muestras induce a dudas sobre qué análisis se han basado en cada una de ellas. Otra dificultad habitual consiste en diferenciar entre los casos en los que “no se aporta información, pero no es fundamental dada la finalidad del test” y aquellos en los que “no se aporta información y esta se considera fundamental”, distinción clave porque en el primer caso la valoración es N/A, mientras que en el segundo debe puntuarse con 0. En este punto, los revisores expertos en el contenido del test tienden a experimentar mayor inseguridad, dada la exigencia de conocimientos psicométricos necesarios para tomar esta decisión. Asimismo, se observan discrepancias entre revisores, que en algunos casos podrían deberse a diferencias en la interpretación de los criterios de valoración. Ante esta situación, la elaboración de una guía práctica con ejemplos concretos podría facilitar el trabajo de los revisores y aumentar la consistencia de las evaluaciones.

Ya se han realizado 12 procesos de evaluación, lo que confirma la solidez del proyecto de evaluación de test en España y su valor como herramienta de apoyo a la práctica profesional. La información recopilada no solo facilita la elección fundamentada de instrumentos, sino que también ofrece a editores y autores indicaciones claras sobre aspectos susceptibles de mejora. De cara al futuro, sería conveniente desarrollar guías prácticas que faciliten la aplicación consistente del CET-R en aquellos aspectos metodológicos que generan mayor dificultad entre los revisores, contribuyendo así a mejorar la consistencia inter-revisores. Asimismo, el desarrollo de una versión digital del CET-R que permita su aplicación en línea permitiría automatizar el cálculo de las puntuaciones finales a partir de las valoraciones parciales, reduciendo la carga de trabajo y los posibles errores asociados al proceso manual de cumplimentación. Finalmente, reforzar la difusión de los resultados y promover una formación continua en psicometría y evaluación de test permitirá que los profesionales en Psicología aprovechen plenamente este recurso y lo integren de manera efectiva en su práctica, contribuyendo a una práctica profesional más rigurosa, responsable y ajustada a las exigencias éticas de la disciplina.

Agradecimientos

En primer lugar, quiero expresar mi agradecimiento a los profesores de las distintas universidades que participaron como revisores, cuya labor ha sido fundamental para el desarrollo de esta evaluación. También agradezco a los miembros de la Comisión de Test, especialmente a Ana Hernández, por su constante apoyo y orientación a lo largo de todo el proceso. Mi reconocimiento se extiende al personal administrativo del COP y a las casas editoriales implicadas. La colaboración y disposición de todos ellos fueron esenciales para la realización de la duodécima evaluación de test editados en España.

Conflicto de Intereses

No hay conflicto de intereses.

Referencias

- Abad, F. J. (2024). Décima evaluación de test editados en España: Incorporando información sobre test no comerciales. *Papeles del Psicólogo*, 45(2), 56-64. <https://doi.org/10.23923/pap.psicol.3033>
- Abad, F. J., Sánchez-Sánchez, F., y Santamaría, P. (2020). *Matrices-TAI: Test adaptativo de inteligencia general*. Hogrefe TEA Ediciones.
- American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (7th ed.). American Educational Research Association. <https://www.apa.org/science/programs/testing/standards>
- Blanco, I., Morón, A., Minaya, N., y Machaca-Luna, R. (2023). *Screening de Bienestar Emocional (SBE)*. Habilmind.
- Cohen, R. J., y Swerdlik, M. E. (2018). *Psychological testing and assessment: An introduction to test and measurement* (9th ed.). New York, NY: McGraw-Hill Education.
- Elosua, P., y Geisinger, K. F. (2016). Cuarta evaluación de test editados en España: Forma y fondo. *Papeles del Psicólogo*, 37(3), 82-88. <https://www.papelesdelpsicologo.es/pdf/2693.pdf>
- Evers, A., Muñiz, J., Hagemester, C., Hstmælingen, A., Lindley, P., Sjöberg, A., y Bartram, D. (2013). Assessing the quality of test: Revision of the EFPA review model. *Psicothema*, 25(3), 283-291. <https://doi.org/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*, 38(3), 161-168. <https://doi.org/10.23923/pap.psicol2017.2844>
- Hernández, A., Tomás, J. M., Ferreres, A., y Lloret, S. (2015). Tercera evaluación de test editados en España. *Papeles del Psicólogo*, 36(1), 1-8. <https://www.papelesdelpsicologo.es/pdf/2484.pdf>
- Hernández, A., Paradell, È., Saavedra, L., y Vallar, F. (2023). *MACI-II: Adaptación española del Millon Adolescent Clinical Inventory-II*. Pearson Educación.
- Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G., y Elosua, P. (2016). Revisión del modelo para evaluar la calidad de los test utilizados en España. *Papeles del Psicólogo*, 37(1), 192-197. <https://www.papelesdelpsicologo.es/pdf/2775.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*, 40(1), 21-30. <https://doi.org/10.23923/pap.psicol2019.2886>
- 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>
- Guilera, G., y Barrios, M. (2025). Undécima evaluación de test editados en España. *Papeles del Psicólogo*, 46(3), 158-166. <https://www.papelesdelpsicologo.es/resumen?pii=3069>
- Lozano, L. M. (2023). Novena evaluación de los test editados en España. *Papeles del Psicólogo*, 44(1), 1-7. <https://doi.org/10.23923/pap.psicol.3004>
- Martínez de la Iglesia, J., Dueñas-Herrero, R., Onís Vilches, M. C., Aguado-Taberné, C., Colomer, C., y Luque-Luque, R. (2001). Adaptación y validación al castellano del cuestionario de Pfeiffer (SPMSQ) para detectar la existencia de deterioro cognitivo en personas mayores de 65

- años. *Medicina Clínica*, 117(4), 129-134. [https://doi.org/10.1016/S0025-7753\(01\)72040-4](https://doi.org/10.1016/S0025-7753(01)72040-4)
- Mayer, J. D., y Salovey, P. (1997). What is emotional intelligence? En P. Salovey y D. J. Sluyter (Eds.), *Emotional development and emotional intelligence: Implications for educators* (pp. 3-31). Basic Books.
- Merchán, I. M., González, J. D., y Ramos, J. L. (2023). *THINEME: Test de Habilidad en Inteligencia Emocional: Manual técnico*. Giunti Psychometrics.
- 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., Hernández, A., y Ponsoda, V. (2015). Quinta evaluación de test editados en España: Mirando hacia atrás, construyendo el futuro. *Papeles del Psicólogo*, 38(3), 161-168. <https://doi.org/10.23923/pap.psicol2017.2844>
- Muñiz, J., Fernández-Hermida, J. R., Fonseca-Pedrero, E., Campillo-Álvarez, Á., y Peña-Suárez, E. (2011). Primera evaluación de test editados en España. *Papeles del Psicólogo*, 32(2), 113-128. <https://www.papelesdelpsicologo.es/pdf/1947.pdf>
- Pfeiffer, E. (1975). A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients. *Journal of the American Geriatrics Society*, 23(10), 433-441. <https://doi.org/10.1111/j.1532-5415.1975.tb00927.x>
- Ponsoda, V., y Hontangas, P. (2013). Segunda evaluación de test editados en España. *Papeles del Psicólogo*, 34(2), 82-90. <https://www.papelesdelpsicologo.es/pdf/2232.pdf>
- Prieto, G., y Muñiz, J. (2000). Un modelo para evaluar la calidad de los test utilizados en España. *Papeles del Psicólogo*, 77, 65-77. <https://www.papelesdelpsicologo.es/resumen?pii=1102>
- Ramos, J. L., y Cuadrado, I. (2006). *PECO - Prueba para la Evaluación del Conocimiento Fonológico*. EOS.
- Ramos, J. L., González, A. I., y Gutiérrez, R. (2023). *TECO - Test de Evaluación de la Conciencia Fonológica*. Giunti Psychometrics.
- Sánchez-Sánchez, F., Santamaría, P., y Abad, F. J. (2015). *Matrices: Test de inteligencia general*. TEA Ediciones.
- Valderrama, B. (2010). *Motivación inteligente*. Madrid: Prentice Hall.
- Valderrama, B. (2018). La rueda de motivos: Hacia una tabla periódica de la motivación humana. *Papeles del Psicólogo*, 39(1), 60-67. <https://doi.org/10.23923/pap.psicol2018.2855>
- Valderrama, B., Escorial, S., y Luceño, L. (2015). *APM. Análisis del Perfil Motivacional*. TEA Ediciones.
- Viladrich, C., Doval, E., Penelo, E., Aliaga, C., Espelt, A., García-Rueda, B., y Angulo-Brunet, A. (2021). Octava evaluación de test editados en España: Una experiencia participativa. *Papeles del Psicólogo*, 42(1), 1-9. <https://doi.org/10.23923/pap.psicol2020.2937>