

Artículo

Decimotercera Evaluación de Test Editados en España

Sergio Escorial Martín 

Universidad Complutense de Madrid, España

INFORMACIÓN

Recibido: Octubre 27, 2025

Aceptado: Enero 14, 2026

Palabras clave

Test
Evaluación test
Psicometría
Propiedades psicométricas
CET-R

RESUMEN

La evaluación de la calidad de los test psicológicos es una labor fundamental para garantizar una práctica profesional ética y eficaz. Este trabajo presenta los resultados de la decimotercera edición del proyecto de evaluación de test editados en España, una iniciativa continua de la Comisión de Test del Consejo General de la Psicología. En esta edición, se analizaron seis instrumentos (cinco comerciales: ATENTO, BFQ-2, EBE, EVT-3, PPVT-5; y uno no comercial: CoPsoQ-Istas21) utilizando el Cuestionario de Evaluación de Test Revisado (CET-R v1.1) y un sistema de revisión por pares expertos. Los resultados revelan una marcada heterogeneidad. Si bien la calidad de los materiales y la fundamentación teórica es generalmente buena, se han apreciado deficiencias críticas en la evidencia psicométrica reportada. Concretamente, se detectaron carencias en algunos apartados de las evidencias de validez. Estos hallazgos subrayan la necesidad de que los profesionales realicen un consumo crítico de las pruebas que van a emplear, y de que las editoriales y los autores refuercen la justificación psicométrica de sus instrumentos.

Thirteenth Review of Tests Published in Spain

ABSTRACT

The assessment of the quality of psychological tests is a fundamental task for guaranteeing ethical and effective professional practice. This paper presents the results of the thirteenth edition of the evaluation project for tests published in Spain, an ongoing initiative by the Test Commission of the Spanish General Council of Psychology. In this edition, six instruments were analyzed (five commercial: ATENTO, BFQ-2, EBE, EVT-3, PPVT-5; and one non-commercial: CoPsoQ-Istas21) using the Revised Test Evaluation Questionnaire (CET-R v1.1) and an expert peer-review system. The results reveal marked heterogeneity. While the quality of the materials and theoretical foundation is generally good, critical deficiencies have been noted in the reported psychometric evidence. Specifically, shortcomings were detected in some sections of the validity evidence. These findings underscore the need for professionals to critically appraise the tests they plan to use, and for publishers and authors to strengthen the psychometric justification of their instruments.

Keywords

Test
Assessing test quality
Psychometrics
Psychometric properties
CET-R

Cómo citar: Escorial, S. (2026). Decimotercera evaluación de test editados en España. *Papeles del Psicólogo/Psychologist Papers*, 47(3), 150-161.

<https://doi.org/10.70478/pap.psicol.2026.47.17>

Autor de correspondencia: Sergio Escorial Martín sergio.escorial@psi.ucm.es 

Este artículo está publicado bajo Licencia Creative Commons 4.0 CC-BY-NC-ND

Los test han sido descritos como “... *la tecnología más sofisticada de la que disponen los psicólogos para ejercer su profesión*” (Muñiz, 2010, p. 57). Encuestas publicadas por Muñiz desde hace más de 25 años (Muñiz y Fernández-Hermida, 2000, 2010 y Muñiz et al. 2020) constatan que son una herramienta consolidada, de uso habitual y clave para la toma de decisiones en todos los ámbitos psicológicos. Su valor radica en su naturaleza estandarizada y en sus propiedades psicométricas conocidas y evaluables. Los profesionales los necesitan porque “(...) *la evaluación rigurosa está en la base de unos diagnósticos precisos, claves a su vez para generar intervenciones eficaces*” (Muñiz, 2018, p. 23). Por lo tanto, sus propiedades psicométricas (i.e. validez y fiabilidad) no son un mero tecnicismo, sino un pilar fundamental para la efectividad de la práctica profesional.

Resulta esencial que los profesionales de la Psicología entiendan que la mera utilización de un test no garantiza una buena práctica profesional. La utilidad de estas herramientas es directamente proporcional a la calidad de sus propiedades psicométricas, a la adecuación de su baremación y al contexto de uso propuesto. El uso de instrumentos con deficiencias metodológicas, falta de validez o fiabilidad o baremos obsoletos, no solo merma la eficacia de la intervención psicológica, sino que puede suponer un riesgo ético significativo al fundamentar decisiones críticas para la vida de las personas en datos manifiestamente erróneos (Prieto y Muñiz, 2000). Además, un test puede ser excelente para un propósito y no adecuado para otros (AERA, APA y NCME, 2014).

Esta preocupación por la calidad y el uso ético de los test fue precisamente un catalizador que ayudó a la psicometría a trascender desde la vertiente más teórica de esta disciplina hasta una vertiente mucho más centrada en proporcionar respuestas en contextos aplicados. Para guiar a desarrolladores y usuarios, se han establecido marcos de referencia concretos. El más destacado es, sin duda, los *Standards for Educational and Psychological Testing* (AERA, APA y NCME, 2014). Paralelamente, organismos como la *International Test Commission* (ITC) han jugado también un papel crucial, publicando directrices específicas —conocidas como *ITC Guidelines*— para abordar retos como la adaptación de tests entre culturas (Hernández et al., 2020; ITC, 2018, Muñiz et al., 2013) o la evaluación mediada por tecnología y en entornos digitales (ITC y Association of Test Publisher, 2022).

Consciente de esta realidad, y en línea con las directrices internacionales promovidas por asociaciones como la *European Federation of Psychologists' Associations* (EFPA), el Consejo General de la Psicología de España (COP) asumió la responsabilidad de ofrecer orientación a los profesionales de la Psicología. A través de su Comisión de Tests, se estableció hace más de una década una iniciativa sistemática y periódica para evaluar la calidad de los tests editados en España. Este proyecto no solo busca informar al psicólogo sobre las garantías de algunas de las herramientas disponibles, sino también fomentar una cultura de calidad entre las editoriales y los autores de las pruebas, incentivando la mejora continua en la adaptación y desarrollo de nuevos instrumentos (Hernández et al., 2022).

Para llevar a cabo esta tarea, la Comisión de Test adoptó el Cuestionario de Evaluación de Test (CET) desarrollado inicialmente por Prieto y Muñiz (2000). Desde entonces, esta herramienta ha sido modificada y revisada para alinearse con las modificaciones en las directrices internacionales y para incorporar avances técnicos

en el proceso de validación de los test, dando lugar al Cuestionario de Evaluación de Test Revisado (CET-R) (Hernández et al., 2016). En años posteriores se ha publicado en la página web de la Comisión de Test una versión del CET-R con algunas modificaciones (CET-R v1.1) que ya fueron descritas en trabajos previos (Abad, 2024; Guillera y Barrios, 2025). En cualquier caso, de manera general, este instrumento se basa en el modelo de evaluación de la EFPA y proporciona un marco estandarizado para la revisión de las características fundamentales de un test que permiten evaluar la calidad de la prueba. El proceso de evaluación desarrollado e implementado en España por la Comisión de Test es colaborativo, implicando no solo a los miembros de la propia Comisión, sino también a revisores expertos externos —tanto en el área sustantiva del test como en psicometría— y a las propias casas editoriales.

Una evolución notable de esta iniciativa, sistematizada desde la décima edición (Abad, 2024), es la inclusión de pruebas no comerciales. Esta decisión estratégica responde a su uso extendido en la práctica profesional (Muñiz et al., 2020) y a que, al carecer de amparo editorial, suelen adolecer de revisiones psicométricas actualizadas. El objetivo es suplir este vacío de información y ofrecer al profesional el mismo análisis crítico que reciben las pruebas comerciales. No obstante, este proceso conlleva retos metodológicos significativos, ya identificados por Abad (2024), como la frecuente ausencia de manuales técnicos completos, la existencia de múltiples versiones de un mismo instrumento o la dificultad para acotar las fuentes bibliográficas necesarias para la evaluación.

Desde el inicio de este proyecto, se han publicado anualmente los resultados de estas evaluaciones, tanto en la página web de la Comisión de Test (<https://www.cop.es/test/>) como en la revista *Papeles del Psicólogo*. Hasta la fecha, han sido objeto de evaluación 116 test. En la *tabla 1* se muestra el listado de las pruebas evaluadas en cada edición. Se informa del año de la edición y coordinador/a de la misma. Además, se especifica con qué versión del CET se realizó el proceso de evaluación. Como se puede observar, las 4 primeras ediciones se desarrollaron empleando el CET original de Prieto y Muñiz (2000). Las cinco siguientes ediciones emplearon el CET-R (Hernández et al., 2016). Desde la 10ª edición hasta la presente se ha empleado la versión CET-R v1.1. El balance general ha sido siempre positivo, y algunos análisis longitudinales de este proceso, como el presentado en la undécima edición, sugiere incluso que esta evaluación continua en el tiempo ha podido tener un efecto positivo tangible en la calidad y en la cantidad de la información psicométrica que las editoriales reportan en los manuales de las pruebas publicadas más recientemente (Guillera y Barrios, 2025).

Siguiendo esta tradición el presente artículo tiene como objetivo principal presentar los resultados de la decimotercera edición de la evaluación de tests. En esta ocasión, se han analizado seis instrumentos, cubriendo un espectro de dominios que van desde la evaluación comprensiva de aspectos vinculados con el lenguaje, hasta la evaluación de las funciones ejecutivas y el TDAH, pasando por la evaluación de riesgos psicosociales, los cinco grandes rasgos de personalidad o una batería completa para la evaluación del bienestar educativo. Se detallarán las puntuaciones obtenidas por estas seis pruebas en las principales dimensiones del instrumento de valoración de la calidad empleado (CET-R v1.1), y también se discutirán las fortalezas y debilidades más destacables observadas en esta decimotercera edición.

Tabla 1

Listado de Test Evaluados en Cada Edición

Edición	Coordinador/a	Versión CET	Nº de Test	Instrumentos Evaluados
2024	S. Escorial	CET-R (V1.1)	6	ATENTO, BFQ-2, CoPsoQ-Istas21, EBE, EVT-3, PPVT-5
2023	P. Recio	CET-R (V1.1)	7	APM, MACI-II, MATRICES-TAI, SBE, SPMSQ-VE, TECO, THINEME
2022	G. Guillera y M. Barrios	CET-R (V1.1)	7	BESS, CTC-R, CTS-2, CUMANIN-2, MASC2, SRP4, TALE
2021	F. J. Abad	CET-R (V1.1)	7	BASC-3, CIT, CPF, CTE, DABS, GDS, PROLEXIA
2020	L. Lozano	CET-R	6	BAYLEY-III, BECOLE-R, CAG, DAS, MacArthur, Raven 2
2019	C. Viladrich	CET-R	11	BADyG/E1-r, BADyG/E2-r, BADyG/E3-r, BADyG/i, BADyG/M-r, BADyG/S-r, BRIEF-P, CELF-5, MCMI-IV, PECO, TONI-4
2018	L. Gómez	CET-R	8	BRIEF-2, BYI-2, DP-3, Factor-g-R, IAES-A, PAIB-1, PAIB-2, PAIB-3
2017	M.D. Hidalgo	CET-R	10	BADyG/E2-r, BADyG/S-r, BAT-7, BDI-FastScreen, BPR, CESPRO, MATRICES, MBMD, Perfil Sensorial-2, Q-PAD
2016	E. Fonseca	CET-R	12	BADyG/E3, CAEPO, EDI-3, EVAPROMES, LAEA, MABC-2, MABC-2-LOC, NEPSY-II, PAIB (2, 3), PRO (1-2, 3), TEMT, WISC-V
2015	P. Elosua	CET	11	ABAS-II, BADyG/M-r, BETA, BSI-18, CECAD, EHPAP, PAIB (1), PECC, SCIP-S, WMS-IV, WPPSI-IV
2014	A. Hernández	CET	11	BCSE, BECOLE, Boehm-3, Boehm-3 Preescolar, CESQT, ECLE, ESQUIZO-Q, IEI, SOC, Trauma, WAIS-IV
2013	V. Ponsoda	CET	12	BAI, BAS-II, BDI-II, CEAM, CompeTEA, ESCOLA, ESPERI, EPV-R, MPR, PAI, RIAS, WNV
2012	J. Muñiz	CET	8	EFAI, EVALUA, IGF, MMPI-2-RF, NEO PI-R, PROLEC-R, WISC-IV, 16PF-5

Nota: El nombre completo correspondiente al acrónimo de cada test se encuentra listado en el [Anexo](#).

Método

Participantes

De manera análoga a ediciones anteriores, la búsqueda de potenciales revisores se realizó con la intención de que cada prueba fuese revisada por un profesional experto en psicometría y por un profesional experto en las variables sustantivas que pretende medir cada test. Para la identificación de estos expertos se combinaron diferentes estrategias: a) se consultaron las listas de revisores publicadas de las once ediciones de la evaluación de test publicadas en *Papeles del Psicólogo* (Abad, 2024; Elosua y Geisinger, 2016; Fonseca y Muñiz, 2017; Gómez-Sánchez, 2019; Guillera y Barrios, 2025; Hernández et al., 2015; Hidalgo y Hernández, 2019; Lozano, 2023; Muñiz et al., 2011; Ponsoda y Hontangas, 2013; Recio-Saboya, 2026; Viladrich et al., 2021); b) se extrajeron datos de profesionales tras consultar en los listados de asociaciones científicas (i.e. la *Asociación Española de Metodología de las Ciencias del Comportamiento* (AEMCCO) o la *Asociación Iberoamericana para la Investigación de las Diferencias Individuales* (AIIDI)); y c) cuando fue necesario se completó el listado de revisores empleado la red de contactos del coordinador de la presente edición.

Inicialmente se contactó con 12 potenciales revisores, dos para cada prueba con el criterio anteriormente establecido por su área de especialización. En esta primera lista se buscó la paridad de sexo y la variabilidad en la distribución geográfica. Tras un primer contacto con las personas de esta primera lista, 4 de ellas rechazaron participar (33,33% de tasa de rechazo). De estas 4 personas, dos pertenecían al panel de expertos en psicometría y dos al panel de expertos en variables sustantivas del test. A partir de aquí, se contactó con otras 4 personas para reemplazar los perfiles de quienes rechazaron inicialmente. Todas las personas contactadas en este segundo momento aceptaron participar en esta edición como revisores. La [tabla 2](#) recoge el panel de revisores definitivo que participaron en la decimotercera edición de la evaluación de test editados en España.

Tabla 2

Panel de Revisores de la Decimotercera Evaluación de Test

Nombre	Filiación
Alicia Franco Martínez	Universidad Autónoma de Madrid (UAM)
David Gallardo Pujol	Universitat de Barcelona (UB)
María Dolores Hidalgo Montesinos	Universidad de Murcia (UMU)
Manuel Ignacio Ibáñez Ribes	Universitat Jaume I (UJI)
Lourdes Luceño Moreno	Universidad Complutense de Madrid (UCM)
Fabía Morales Vives	Universitat Rovira i Virgili (URV)
David Navarro González	Universitat de Lleida (ULI)
Ricardo Olmos Albacete	Universidad Autónoma de Madrid (UAM)
Jesús Privado Zamorano	C.E.S. Cardenal Cisneros
Patricia Recio Saboya	Universidad Nacional de Educación a Distancia (UNED)
José María Ruiz Sánchez de León	Universidad Complutense de Madrid (UCM)
Miguel Sorrel Luján	Universidad Autónoma de Madrid (UAM)

Instrumento

Para esta decimotercera edición se empleó el Cuestionario de Evaluación de Test Revisado (CET-R v1.1), la versión del protocolo estandarizado utilizada desde la 10ª edición. Al igual que el CET-R (Hernández et al., 2016) esta versión se estructura en tres secciones principales. La primera, *Descripción general del test*, recoge información fáctica y descriptiva sobre la prueba, como los datos editoriales, los constructos que mide, la población diana, el formato de los ítems, los tiempos de aplicación o el precio de los materiales. El núcleo de la evaluación se concentra en la segunda sección, *Valoración de las características del test*, que permite un análisis detallado de cuatro pilares psicométricos: 1) Fundamentación y desarrollo, 2) Validez, 3) Fiabilidad y 4) Baremos e interpretación de puntuaciones. Finalmente, la tercera sección, *Valoración global*, ofrece un resumen cuantitativo de las puntuaciones y un apartado cualitativo abierto para detallar las fortalezas, debilidades y recomendaciones de uso profesional de la prueba.

Los ítems adoptan generalmente el siguiente sistema de etiquetas: N/A = *No aplicable*, 0 = *No se aporta información en la documentación*, 1 = *Inadecuada*, 2 = *Adecuada, pero con algunas carencias*, 3 = *Adecuada*, 4 = *Buena*, y 5 = *Excelente*. La categoría “Excelente” se acompaña de una descripción detallada para guiar a los evaluadores. Además, en aquellos ítems donde es posible una cuantificación más objetiva (i.e. tamaños de las muestras, valor de los coeficientes de fiabilidad, etc.) se utilizan etiquetas específicas para facilitar una valoración más precisa. Por último, los apartados generales que resumen las características globales del test siguen esta misma escala de categorías.

Es importante resaltar que esta versión refina la evaluación mediante tres modificaciones clave: 1) una distinción clara entre información fundamental (cuya ausencia penaliza con un cero) y no crítica (sin impacto en la puntuación); 2) criterios específicos para test adaptados, exigiendo mayor peso para las evidencias psicométricas obtenidas con muestras locales; y 3) pautas técnicas para la correcta valoración del AUC (curvas ROC) en test diagnósticos. Todas las versiones del CET están disponibles en la página web de la Comisión de Test (<https://www.cop.es/test/>).

Procedimiento

Las casas editoriales (Giunti Psychometrics, Pearson Educación y TEA-Hogrefe) propusieron a la Comisión de Test las cinco pruebas comerciales que formarían parte de esta decimotercera edición de evaluación de test editados en España. Además, al igual que en las tres últimas ediciones, la Comisión de Test añadió al listado de pruebas un instrumento no comercial, la versión 2 CoPsoQ-Istas21 adaptada en España (Moncada et al., 2014a, 2014b). La [tabla 3](#) presenta la información de los test evaluados en la presente edición.

Para las cinco pruebas comerciales, y una vez establecido el panel de revisores definitivo, el COP se encargó de gestionar el envío de los materiales a cada revisor. Este envío tuvo lugar durante los meses de septiembre y octubre de 2024. En paralelo, el coordinador enviaba por correo electrónico a cada revisor el CET-R v1.1, estableciendo un plazo de cuatro meses para su

complimentación. Todos los cuestionarios CET-R v1.1 de los revisores se recibieron hacia mediados de febrero de 2025. Entre marzo y mayo de 2025, el coordinador preparó un informe provisional para cada test integrando las valoraciones de ambos revisores. Hay que destacar que no se observaron grandes discrepancias entre las calificaciones de los dos revisores en un mismo test, pero cuando estas aparecieron, el coordinador resolvió la valoración final atendiendo al razonamiento efectuado por los revisores y a la información presente en el manual del test. En mayo de 2025 el coordinador presentó a la Comisión de Test del COP los resultados provisionales de los informes y del proceso hasta esa fecha. En junio de 2025 los informes provisionales de cada prueba fueron enviados a las casas editoriales para que en el plazo de un mes presentaran las alegaciones oportunas. Por último, el coordinador preparó la versión definitiva del informe, considerando la información facilitada por los revisores y por las editoriales en sus alegaciones. Los informes definitivos fueron revisados por un miembro de la Comisión de Test durante el mes de julio de 2025, cuyas sugerencias, fundamentalmente de estilo, fueron incorporadas a los informes. Todos los informes de esta edición se publicaron en la página web de la Comisión de Test en el mes de septiembre de 2025 (<https://www.cop.es/test/>). La [figura 1](#) representa el procedimiento descrito.

Protocolo Específico Para la Escala no Comercial CoPsoQ-Istas21

El Cuestionario Psicosocial de Copenhague (CoPsoQ-Istas21) es una herramienta diseñada para evaluar los principales factores de riesgo psicosocial en entornos laborales. Desde su desarrollo inicial en Dinamarca a finales de los años 90, CoPsoQ ha sido adaptado y validado en más de 25 lenguas, utilizado en cientos de investigaciones y en miles de procesos de evaluación de riesgos psicosociales en el lugar de trabajo. Se ha convertido en una de las herramientas de evaluación más utilizadas y es citado por la OMS como un método de referencia (Leka, 2010), mostrado como un ejemplo de buena práctica por la Agencia Europea para la Seguridad y la Salud en el Trabajo y la Organización Internacional del Trabajo

Tabla 3
Listado de Test Evaluados en la Decimotercera Edición

Acronimo	Nombre	Autores originales (año de publicación)	Autores adaptación (año de publicación)	Editorial
ATENTO	Cuestionario para la Evaluación de las Funciones Ejecutivas y el TDAH	Fernando Sánchez-Sánchez y María Solar (2024)	-	Hogrefe TEA Ediciones
BFQ-2	Big Five Questionnaire - 2	Gian Vittorio Caprara, Claudio Barbaranelli, Laura Borgogni y Michele Vecchione (2007)	Departamento R&D de Giunti Psychometrics España (2021)	Giunti Psychometrics
CoPsoQ-Istas21-V2	Cuestionario Psicosocial de Copenhague Versión 2	Tage S Kristensen, Harald Hannerz, Annie Høgh y Vilhelm Borg (2005)	Salvador Moncada, Clara Llorens, Neus Moreno y Emilia Molinero (2014a)	-
EBE	Batería de Evaluación del Bienestar Educativo	Jesús de la Fuente Arias y José Manuel Martínez-Vicente (2024)	-	Giunti Psychometrics
EVT-3	Expressive Vocabulary Test, Third Edition [Test de vocabulario expresivo-3]	Kathleen T. Williams (2019)	Departamento de I+D de Pearson Clinical Assessment: Ana Hernández, Èrica Paradell, Laura Saavedra y Frédérique Vallar (2024b)	Pearson Educación
PPVT-5	Peabody Picture Vocabulary Test, Fifth Edition [Test de vocabulario en imágenes Peabody-5]	Douglas M. Dunn (2019)	Departamento de I+D de Pearson Clinical Assessment: Ana Hernández, Èrica Paradell, Laura Saavedra y Frédérique Vallar (2024a)	Pearson Educación

(ILO, 2016) y utilizado en informes de la Fundación Europea para la Mejora de las Condiciones de Vida y del Trabajo (Schütte et al., 2014).

En España existen 3 versiones del CoPsoQ-Istas21. La primera versión del CoPsoQ-Istas21 fue adaptada al español con la colaboración del primer autor de la versión original danesa (Moncada et al., 2005). La segunda versión incluía modificaciones sustanciales en el número y la agrupación de las escalas (Moncada et al., 2014b). Finalmente, la tercera versión del CoPsoQ-Istas21 también ha sido adaptada al español por Moncada et al. (2021). Además, de cada una de las versiones existe una forma corta, para empresas con menos de 25 trabajadores; una versión media para entidades con más de 25 trabajadores y una versión larga para ser utilizada en investigación. La presente evaluación ha debido centrarse en la Versión 2 (forma media), puesto que es la única que actualmente ofrece tanto el material documental (manual técnico) como la herramienta informática necesaria para la presente evaluación. Ambos recursos están disponibles en la web del Instituto Sindical de Trabajo, Ambiente y Salud (ISTAS) (<https://copsoq.istas21.net/>).

Debido a que el Manual Técnico del CoPsoQ-Istas21 (Moncada et al., 2014a) carece de los datos psicométricos necesarios para la revisión, remitiendo a Moncada et al. (2014b), se implementó un protocolo de búsqueda bibliográfica sistemática en Scopus y WoS. Se identificaron inicialmente 91 trabajos, que tras un primer cribado se redujeron a 35 documentos potenciales. Para la selección final de los 12 estudios clave, se aplicó un criterio de inclusión basado en la relevancia teórica para la fundamentación del constructo y la

aportación de evidencias empíricas de validez y fiabilidad. Este corpus documental quedó conformado por 4 artículos fundamentales para comprender el desarrollo de la prueba (Kristensen et al., 2005; Moncada et al., 2005, 2014b y 2021) y 8 artículos seleccionados por aportar evidencia psicométrica específica (Delgado-Fernández et al., 2021; García-Iglesias et al., 2021; García-Rodríguez et al., 2015; Louzán et al., 2024; Moncada et al., 2008; Navarro et al., 2022; Useche et al., 2019; Utzet et al., 2014).

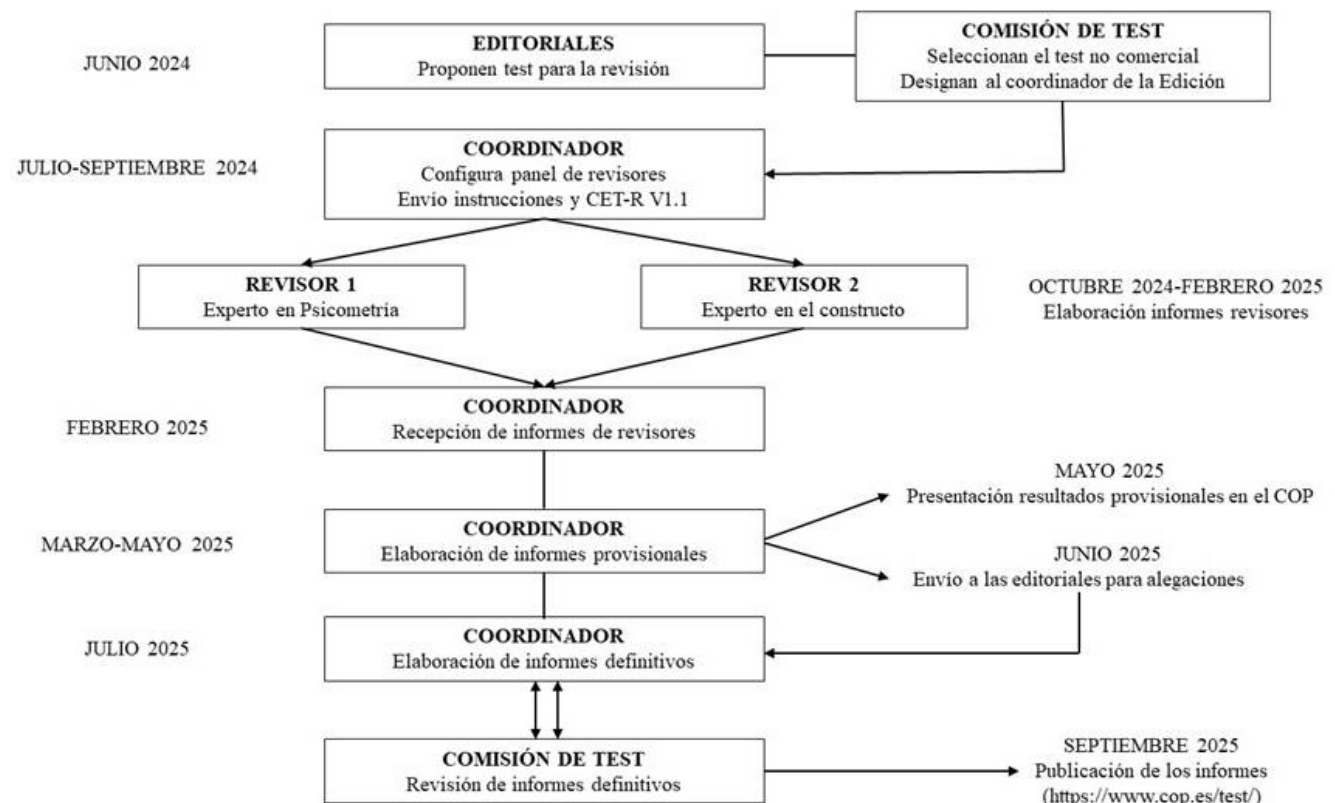
Para garantizar la estandarización del proceso, se facilitó a los revisores un dossier técnico compuesto por: los materiales originales editados por ISTAS (manual y software), el texto completo de los 12 artículos seleccionados y el listado bibliográfico de los 35 trabajos preseleccionados para consulta complementaria. Una vez compilada esta documentación, el proceso de valoración fue idéntico al seguido con los test comerciales (ver Figura 1).

Análisis de Datos

En primer lugar, para evaluar la fiabilidad del proceso de revisión llevado a cabo en esta edición, se estimó el grado de acuerdo entre los evaluadores independientes mediante el Coeficiente de Correlación Intraclass (ICC). Dado el diseño anidado del estudio, en el que los pares de revisores fueron independientes y distintos para cada instrumento evaluado, se especificó un modelo de efectos aleatorios de un factor (*one-way random effects model*), analizando la concordancia en las puntuaciones individuales. A diferencia de aproximaciones metodológicas previas en esta serie de evaluaciones de test editados en España, que reportaron la mediana de las

Figura 1

Representación del Procedimiento de la Decimotercera Edición de la Evaluación de Test Editados en España



correlaciones de Pearson para cada prueba evaluada obtenida con las respuestas de los dos revisores (i.e., [Abad, 2024](#); [Ponsoda y Hontangas, 2013](#)), en el presente estudio se optó por realizar el cálculo sobre el conjunto de datos. Esta decisión metodológica se fundamenta en la necesidad de evitar el sesgo de restricción de rango que ocurre al analizar instrumentos aislados con alta homogeneidad interna, lo cual tiende a subestimar artificialmente la fiabilidad. El análisis global permite capturar la varianza completa a través de los distintos niveles de calidad de los test, ofreciendo una estimación más robusta de la consistencia del proceso de evaluación.

A continuación, se analizaron los resultados generales obtenidos en la segunda sección (cuantitativa) del CET-R v1.1. Se calcularon los valores medios de cada uno de los seis test evaluados en los apartados que componen los bloques de esta sección (Fundamentación y desarrollo, Validez, Fiabilidad y Baremos e interpretación de puntuaciones). Para contextualizar los resultados de la presente edición, las puntuaciones promedias obtenidas por los seis instrumentos evaluados en estos apartados (Promedio Edición), se contrastaron con los promedios históricos acumulados. Estas comparaciones se realizaron a nivel puramente descriptivo, analizando la posición de las puntuaciones promedio observadas en esta edición con respecto a la tendencia central de dos series históricas: Histórico 1: promedios calculados a partir de las pruebas de la 10ª, 11ª y 12ª edición, que están basadas en el CET-R v1.1; e Histórico 2: promedios calculados a partir de las 9 ediciones anteriores, las cuatro primeras basadas en el CET y las cinco siguientes en el CET-R. Para cada apartado solamente se tuvieron en cuenta los datos de las pruebas que presentaban puntuación, que no eran los mismos en el CET y en el CET-R. (serie agregada acumulada). Por estas razones, dado el carácter heterogéneo de los datos históricos, no se aplicaron pruebas de inferencia estadística formal para contrastar diferencias de medias. No obstante, para

facilitar una interpretación sustantiva se calcularon las desviaciones típicas y los intervalos de confianza al 95%, tanto de los datos de la presente edición como de las dos series históricas. En la interpretación de los resultados, se consideró que existen diferencias dignas de mención cuando la puntuación media se sitúa fuera de los límites del intervalo de confianza del promedio histórico de referencia, advirtiendo siempre sobre la naturaleza descriptiva de estas apreciaciones.

Resultados

Los informes definitivos correspondientes a los 6 test evaluados en esta decimotercera edición están disponibles para ser consultados desde la página de la Comisión de Test, dentro del apartado correspondiente al año 2024 (<https://www.cop.es/test/>).

En primer lugar, el análisis de la fiabilidad inter-jueces se llevó a cabo sobre un total de $N = 160$ pares de observaciones válidas, correspondientes a los ítems cuantitativos del CET-R v1.1 aplicables a los seis instrumentos. Los resultados arrojaron un Coeficiente de Correlación Intraclass (ICC) de ,84, con un intervalo de confianza al 95% situado entre ,79 y ,88. De acuerdo con los criterios de interpretación establecidos por [Cicchetti \(1994\)](#) o [Koo y Li \(2016\)](#), este valor indica un nivel de acuerdo excelente entre los revisores independientes. Este hallazgo supera sustancialmente las estimaciones de fiabilidad reportadas en las ediciones precedentes (con medianas situadas en torno a ,60), lo que sugiere que la variabilidad observada en las puntuaciones es atribuible fundamentalmente a las características psicométricas de los test evaluados y no a los evaluadores, respaldando así la solidez técnica de los resultados presentados en esta edición.

En la [tabla 4](#) se presenta un resumen de las puntuaciones obtenidas en cada uno de los apartados generales de la segunda sección del

Tabla 4
Puntuaciones Obtenidas por los Test Evaluados en la Decimotercera Edición

	ATENTO	BFQ-2	CoPsoQ-Istas21	EBE	EVT-3	PPVT-5	Promedio Edición	Histórico ¹	Histórico ²
Desarrollo									
Materiales y documentación	5	3,5	2	4	5	5	4,1 (1,20) [2,8-5,0]	4,3 (,67) [4,0-4,6]	4,3 (,62) [4,2-4,4]
Fundamentación teórica	5	4	4	4	3,5	3,5	4,0 (,55) [3,4-4,6]	3,9 (1,01) [3,4-4,4]	4,1 (,82) [4,0-4,3]
Adaptación	-	2	2	-	4,5	4,5	3,3 (1,44) [1,0-5,0]	3,9 (,70) [3,4-4,4]	4,3 (,70) [4,1-4,5]
Análisis de ítems	5	3	2	3	3	3	3,2 (,98) [2,1-4,2]	3,7 (,88) [3,3-4,2]	3,8 (,73) [3,7-4,0]
Validez									
Contenido	5	2,5	1	3	3	3	2,9 (1,28) [1,6-4,3]	3,6 (1,04) [3,1-4,1]	3,8 (,79) [3,6-4,0]
Relación con otras variables	5	3,5	3	2,5	3,8	3,7	3,6 (,85) [2,7-4,5]	3,5 (1,02) [3,0-4,0]	3,5 (,64) [3,4-3,7]
Estructura interna	5	2	1	2,5	0	0	1,8 (1,89) [0-3,7]	3,4 (1,12) [2,8-3,9]	3,8 (,86) [3,6-4,0]
Análisis del DIF	5	3	0	3	3	3	2,8 (1,60) [1,2-4,5]	4,1 (,82) [3,5-4,7]	4,0 (1,02) [3,4-4,5]
Fiabilidad									
Equivalencia	-	-	-	-	-	-	-	-	3,6 (,66) [2,9-4,3]
Consistencia interna	5	4	3	3,5	4	4	3,9 (,66) [3,2-4,6]	3,9 (,79) [3,6-4,3]	4,2 (,67) [4,1-4,4]
Estabilidad	4,5	3	-	-	3,5	3,5	3,6 (,63) [2,6-4,6]	3,3 (,64) [3,0-3,7]	3,5 (,78) [3,2-3,7]
TRI	-	-	-	-	-	-	-	4,0 (,71) [3,3-4,7]	4,5 (,45) [4,0-5,0]
Inter-jueces	4	-	-	-	-	-	4,0	3,9 (1,31) [1,8-5,0]	4,5 (,84) [3,6-5,0]
Baremos e interpretación de puntuaciones	5	3	3,3	3,3	4,3	4,3	3,9 (,78) [3,1-4,7]	3,8 (,78) [3,5-4,2]	4,1 (,68) [3,9-4,2]
Total	4,9	3,1	2,1	3,2	3,4	3,4			

Nota: Entre paréntesis las desviaciones típicas (DT) y entre corchetes los IC al 95%. Las comparaciones con la columna Histórico 1 son directas al emplear la misma versión del instrumento (CET-R v1.1). Los estadísticos descriptivos de la columna Histórico 2 (especialmente DT e IC) deben interpretarse con cautela como una aproximación referencial histórica, asumiendo cierta heterogeneidad métrica derivada de la integración de datos procedentes de las dos versiones anteriores del instrumento (CET y CET-R) en aquellas dimensiones conceptualmente equivalentes.

CET-R v1.1. Con las valoraciones de los seis test en cada apartado se ha calculado un promedio (Promedio Edición) y, para poder establecer comparaciones meramente descriptivas se han calculado dos promedios históricos. El primero de ellos correspondiente a la columna Histórico¹ se ha obtenido con las valoraciones de las 21 pruebas analizadas en las tres ediciones anteriores (ediciones 10^a a 12^a) ya que en estas se empleó la misma versión del CET-R v1.1. El segundo de ellos correspondiente a la columna Histórico² se ha obtenido con las puntuaciones de las 89 pruebas analizadas en las nueve primeras ediciones (ediciones 1^a a 9^a). Entre paréntesis se presenta la desviación típica estimada y entre corchetes los IC 95% para la media estimada. Como se puede observar, en algunos apartados se aprecian discrepancias importantes con el patrón histórico obtenido de las ediciones anteriores.

Resultados de los Test Evaluados en la 13^a Edición

Los resultados obtenidos por los seis instrumentos analizados en la presente edición muestran una marcada heterogeneidad en las puntuaciones, tanto entre las distintas dimensiones del modelo CET-R v1.1, como entre los propios test evaluados (ver [Tabla 4](#)).

En lo referente al bloque de desarrollo, la calidad de los materiales y la documentación es el aspecto mejor valorado, con una media global de 4,1. Instrumentos como el ATENTO, EVT-3 y PPVT-5 alcanzaron la puntuación máxima (5) en este apartado, destacando por la claridad y calidad de sus manuales. Asimismo, la fundamentación teórica se mantiene en niveles elevados (media de 4,0), indicando que el marco conceptual de las pruebas está, por lo general, bien definido. Sin embargo, se observan debilidades en algunos de los procedimientos de adaptación (media de 3,3) y en el análisis de ítems (media de 3,2), donde pruebas como el CoPsoQ-Istas21 presentan puntuaciones más discretas.

El bloque de evidencias de validez es donde se registran las mayores discrepancias y las puntuaciones promedio más bajas. Si bien la relación con otras variables alcanza niveles aceptables (media de 3,6), la validez de contenido (media de 2,9) y, muy especialmente, la validez de estructura interna (media de 1,8), reflejan carencias significativas en la información aportada. Es importante destacar que este bajo promedio en estructura interna se debe a la ausencia de estudios actualizados o específicos en varios de los instrumentos analizados (ej. EVT-3 y PPVT-5 obtuvieron puntuaciones de 0 en este criterio específico bajo los estándares del CET-R v1.1), contrastando con la excelencia mostrada por el test ATENTO (5).

En cuanto a la fiabilidad, los resultados son generalmente satisfactorios. La consistencia interna muestra una solidez transversal a la mayoría de las pruebas (media de 3,9), mientras que la estabilidad temporal (media de 3,6) presenta un mayor IC dependiendo del diseño empleado en la obtención para esta evidencia en cada instrumento. En este bloque de fiabilidad destaca el ATENTO, no solo por las elevadas puntuaciones que obtiene en los apartados, sino también por ser la única prueba de las evaluadas en esta edición que presenta evidencias de fiabilidad inter-jueces. Finalmente, la baremación e interpretación de puntuaciones alcanza un nivel bueno (3,9), con la mayoría de los test aportando normas adecuadas y actualizadas. En este apartado, ATENTO, EVT-3 y PPVT-5 muestran una calidad a destacar. Hay que resaltar que estas pruebas han elaborado sus baremos empleando procedimientos de tipificación continua que optimizan la eficiencia del proceso de elaboración de

baremos en constructos que siguen una tendencia a través de la edad de los sujetos (Evers et al., 2010; Evers et al., 2013).

Analizando las pruebas en su conjunto, las pruebas comerciales obtienen puntuaciones totales más elevadas que la prueba no comercial, el CoPsoQ-Istas21, que alcanza una puntuación global de 2,1. Destaca la puntuación global del ATENTO de 4,9, mientras que las valoraciones medias del resto de las pruebas comerciales estuvieron por encima del “Adecuado” siendo 3,4; 3,4; 3,2 y 3,1 para EVT-3, PPVT-5, EBE y BFQ-2 respectivamente.

Resultados Comparativos del Promedio de Esta Edición con las Series Históricas

Al contrastar los resultados actuales con los promedios históricos —tanto la serie reciente bajo el mismo estándar (*Histórico 1*) como la serie acumulada con las versiones anteriores del CET (*Histórico 2*)—, se observan patrones divergentes que merecen una lectura detallada. Cabe señalar que estas comparaciones se realizan a nivel descriptivo, considerando los intervalos de dispersión presentados en la [Tabla 4](#).

En las dimensiones de materiales y documentación y de fundamentación teórica, la presente edición se alinea con la tendencia histórica. Los promedios actuales son muy similares a los promedios en *Histórico 1* y en *Histórico 2*, situándose dentro de los márgenes de variabilidad esperados. Esto sugiere que la calidad formal y teórica de la edición de test en España mantiene un estándar de calidad consolidado en estos apartados. Sin embargo, en los apartados de adaptación y de análisis de ítems, los promedios de esta edición parecen ser inferiores, en términos descriptivos, a los promedios de las dos series históricas consideradas.

Por otra parte, se detecta un descenso sustancial en las evidencias de validez respecto a lo observado en ediciones anteriores. La puntuación media en validez de estructura interna de esta edición es, desde una perspectiva descriptiva, inferior tanto al *Histórico 1* como al *Histórico 2*. De igual forma, la validez de contenido se sitúa por debajo descriptivamente de los referentes históricos. En el apartado de análisis de DIF también se observa una puntuación más baja en esta edición a nivel descriptivo. Este hallazgo no implica necesariamente una peor calidad intrínseca de las herramientas, pero sí alerta sobre una disminución en la exhaustividad de los estudios psicométricos reportados en los manuales de esta edición donde se han obtenido puntuaciones de 0 y 1 por varias pruebas y en varios apartados (ver [tabla 4](#)).

Finalmente, en el ámbito de la fiabilidad y baremación, los datos de esta 13^a edición son consistentes con la trayectoria de ambas series históricas. La fiabilidad por consistencia interna replica exactamente el promedio de la serie más reciente (*Histórico 1*) y se mantiene competitiva frente al histórico anterior (*Histórico 2*). De manera similar, la calidad de los baremos no presenta diferencias relevantes a nivel descriptivo respecto al promedio histórico reciente, evidenciando que la estandarización normativa sigue siendo una prioridad en la publicación de test en nuestro contexto.

Conclusiones

El objetivo principal de este trabajo ha sido presentar los resultados de la decimotercera edición de la evaluación de test editados en España, continuando una labor que la Comisión de Test

del Consejo General de la Psicología (COP) viene realizando desde hace más de una década. En esta edición se han analizado 6 instrumentos: ATENTO, BFQ-2, CoPsoQ-Istas21, EBE, EVT-3 y PPVT-5, cuyos resultados globales reflejan una notable heterogeneidad en la calidad de la evidencia reportada. El hallazgo más significativo de esta decimotercera evaluación es la constatación de una marcada dicotomía: mientras que los aspectos formales y de presentación mantienen un nivel bueno de calidad, han aparecido deficiencias críticas en apartados psicométricos nucleares, especialmente en la evidencia de validez de estructura interna y en los análisis de DIF.

En el lado positivo, los resultados de esta edición son coherentes con los de evaluaciones previas (p.ej., [Abad, 2024](#); [Guillera y Barrios, 2025](#)) al mostrar que las casas editoriales cuidan mucho la presentación del producto. Las puntuaciones promedio en "Materiales y documentación" (4,1) y "Fundamentación teórica" (4,0) fueron contundentes en este sentido. Asimismo, la calidad de la baremación (3,9) y los estudios de fiabilidad (3,9) obtuvieron también valoraciones buenas en línea con los promedios históricos de las anteriores ediciones. Esto sugiere que el mercado español ofrece, en su mayoría, herramientas bien documentadas, con manuales claros y fundamentados, y con baremos actualizados y bien elaborados, un aspecto crucial para la correcta interpretación de las puntuaciones de acuerdo con los estándares internacionales ([AERA, APA y NCME, 2014](#)).

Sin embargo, este panorama positivo se ve ensombrecido por la aparición de ciertas carencias muy significativas en la evidencia psicométrica que sustenta la validez de las inferencias. El promedio de la puntuación en "Validez de estructura interna" (1,8) es notablemente bajo, especialmente cuando se compara con los promedios históricos de las ediciones anteriores. Este dato indica que una proporción importante de los manuales evaluados en esta edición no reportaron los análisis factoriales necesarios para demostrar que la estructura del test se corresponde con el modelo teórico que dice medir. Las propias casas editoriales tampoco presentaron estos análisis como evidencias adicionales durante la fase de alegaciones.

La deficiencia detectada en la evidencia de la estructura interna supone una amenaza directa al núcleo de la validez del test. Esta es una fuente de evidencia indispensable para la validación de constructo ([AERA, APA y NCME, 2014](#)), una exigencia inflexible que se aplica con el mismo rigor a los test que postulan una única dimensión. La unidimensionalidad no es una asunción, sino una hipótesis explícita que debe ser demostrada empíricamente ([Brown, 2015](#)), requiriendo como mínimo un Análisis Factorial Confirmatorio (CFA) que contraste el modelo postulado frente a alternativas plausibles. Asumir la unidimensionalidad sin probarla, ignorando una posible multidimensionalidad latente, es un error metodológico: la puntuación total se convierte en una amalgama ininterpretable ([Prieto y Muñiz, 2000](#)) que invalida las inferencias sobre el rasgo ([Reise et al., 2013](#)). Sin una base empírica sólida de la estructura interna, toda evidencia posterior queda comprometida, pues es logísticamente imposible saber con qué se correlaciona una puntuación si se desconoce qué es lo que esta representa.

Algo similar se observó para el promedio obtenido en el análisis del Funcionamiento Diferencial de los Ítems (DIF), que en esta edición fue de 2,8. Los *Standards* ([AERA, APA y NCME, 2014](#)) son explícitos sobre la responsabilidad de los desarrolladores de

aportar evidencia de equidad y demostrar que los ítems no funcionan de manera distinta para subgrupos poblacionales relevantes (p.ej., por género). La falta de estudios DIF supone una amenaza directa a la validez y la justicia de las evaluaciones, un aspecto ético que no puede ser ignorado ([Abad et al., 2011](#); [Escorial, 2017](#); [Escorial y Navas, 2007](#)).

Es fundamental señalar que los promedios globales ocultan una marcada variabilidad, tanto entre los propios test como entre sus distintos apartados. En esta decimotercera edición, coexisten instrumentos como el ATENTO, que ejemplifica las buenas prácticas obteniendo la máxima puntuación (5) en la mayoría de apartados. En el extremo opuesto se encuentra la prueba no comercial, CoPsoQ-Istas21, con un promedio general de 2,1. Esta disparidad evidencia que factores externos como la novedad de la publicación o el prestigio de la institución editora no constituyen, por sí mismos, garantías suficientes de la bondad psicométrica de un instrumento, lo que hace ineludible una revisión crítica de las evidencias de validez y fiabilidad específicas de cada test antes de su uso.

Las implicaciones prácticas son directas: un manual de alta calidad formal no garantiza la solidez psicométrica. El profesional tiene la responsabilidad ética de analizar críticamente las evidencias de validez, fiabilidad y baremos antes de su uso. Estas evaluaciones anuales no constituyen un *ranking*, sino una guía de consumo crítico que expone tanto fortalezas como, crucialmente, las ausencias de información relevante en los manuales. La utilidad del proyecto es validada por los propios psicólogos colegiados. Los datos indican que quienes conocen las evaluaciones las valoran positivamente y las consideran una herramienta necesaria que influye en su selección de instrumentos ([Muñiz et al., 2020](#)). No obstante, esta valoración contrasta con el bajo conocimiento general del proyecto, pues la misma encuesta reveló que solo un 22,5% de los psicólogos españoles conocían su existencia. Este dato señala la necesidad de reforzar las estrategias de difusión por parte de la Comisión de Test.

Por otro lado, la evaluación de pruebas no comerciales debe mantenerse y potenciarse. Aunque estos instrumentos de dominio público muestren, a menudo, deficiencias psicométricas mayores y obtengan peores valoraciones que las pruebas comerciales, persistir en su evaluación es una decisión ineludible. La justificación es doble: primero, su uso frecuente en la práctica profesional ([Muñiz et al., 2020](#)), y segundo, su carencia de amparo editorial. En este sentido, el hecho de que la baja puntuación del CoPsoQ-Istas21 pueda reflejar una ausencia de documentación y no necesariamente una baja calidad psicométrica podría abrir la reflexión sobre la conveniencia de desarrollar en el futuro un protocolo específico para la evaluación de pruebas no comerciales. La Comisión Nacional de Test podrá debatir sobre esta cuestión y tomar la decisión que estime más pertinente. Sin embargo, bajo mi criterio personal, que un test sea "gratuito" no lo exime de demostrar su calidad bajo los mismos estándares que el resto de pruebas empleadas por los profesionales de la Psicología en diferentes ámbitos ([AERA, APA y NCME, 2014](#)). Por tanto, la Comisión de Test hace un ejercicio de responsabilidad al ofrecer este análisis crítico de estas pruebas que protege al usuario final.

Finalmente, este trabajo presenta limitaciones. El CET-R evalúa la calidad de la información reportada en el manual. No hace una re-evaluación psicométrica de los instrumentos. Por tanto, una puntuación baja puede significar, como señalamos anteriormente,

que la evidencia no existe o, simplemente, que no fue incluida en la documentación aportada. Asimismo, el proyecto cubre un número reducido de test anualmente, sin abarcar la totalidad del mercado. El instrumento empleado (CET-R v1.1), aunque útil y sencillo, está llamado a ser modificado próximamente por las nuevas directrices europeas. Con independencia de estos cambios, informatizar el CET-R aportaría ventajas futuras, como minimizar inconsistencias en las puntuaciones o permitir la ponderación de los ítems dentro de cada apartado.

En conclusión, la decimotercera edición de la evaluación de test reitera la utilidad y pertinencia de este proyecto de la Comisión de Tests. Si bien es cierto que se han observado ejemplos de excelencia en el desarrollo de pruebas, persisten lagunas importantes en las evidencias de validez, que rompen la tendencia observada en las últimas ediciones. Se insta a las editoriales y a los propios autores a reforzar estos apartados psicométricos fundamentales, y a los profesionales a ejercer un consumo crítico y fundamentado de las herramientas que, en última instancia, impactan en el bienestar de las personas a las que evalúan.

Agradecimientos

Agradezco a la Comisión de Test su compromiso en esta iniciativa, especialmente a Ana Hernández por su asistencia durante esta edición; a Arantxa Sánchez (COP) por la labor logística; y a las casas editoriales, tanto por los materiales como por su constructivo feedback. Finalmente, expreso mi más sincero agradecimiento a los revisores de esta edición: su minucioso y generoso trabajo fue un aspecto crítico para este proyecto.

Conflicto de Intereses

No existe 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.
- Abad, F. J., Olea, J., Ponsoda, V., y García, C. (2011). *Medición en ciencias sociales y de la salud*. Ed. Síntesis.
- American Educational Research Association, American Psychological Association y National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Psychological Association.
- Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2ª ed.). The Guilford Press.
- Caprara, G. V., Barbaranelli, C., Borgogni, L., y Vecchione, M. (2007). *Big Five Questionnaire-2: Manuale*. Giunti OS: Organizzazioni Speciali.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284-290.
- de la Fuente, J., y Martínez-Vicente, J. M. (2024). *Batería para la Evaluación del Bienestar Educativo*. Giunti Psychometrics.
- Delgado-Fernández, V., Rey-Merchán, M. D. C., y Arquillos, A. L. (2021). Estudio comparativo de los riesgos psicosociales laborales entre profesionales médicos. *Revista de la Asociación Española de Especialistas en Medicina del Trabajo*, 30(1), 24-33.
- Departamento R&D de Giunti Psychometrics España. (2021). *BFQ-2, cuestionario "Big Five" de personalidad-2: Manual* [Adaptación española]. Giunti Psychometrics.
- Dunn, D. M. (2019). *Peabody Picture Vocabulary Test: PPVT-5* (5th ed.). NCS Pearson
- Elosua, P., y Geisinger, K. F. (2016). Cuarta evaluación de tests editados en España: Forma y fondo. *Papeles del Psicólogo*, 37(2), 82-88.
- Escorial, S. (2017). Análisis de la variable sexo en la escala de Búsqueda de Sensaciones (SSS-V) empleando técnicas de Funcionamiento Diferencial de los Ítems. *Avances en Psicología Latinoamericana*, 35(2), 387-405.
- Escorial, S., y Navas, M. J. (2007). Analysis of the gender variable in the Eysenck Personality Questionnaire-Revised scales using differential item functioning techniques. *Educational and Psychological Measurement*, 67, 990-1001.
- Evers, A., Sijtsma, K., Lucassen, W., y 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.
- Evers, A., Muñiz, J., Hagemester, C., Høstmælingen, A., Lindley, P., Sjöberg, A., y Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema*, 25(3), 283-291.
- Fonseca, E., y Muñiz, J. (2017). Quinta evaluación de tests editados en España: mirando hacia atrás, construyendo el futuro. *Papeles del Psicólogo*, 38(3), 161-168.
- García-Iglesias, J. J., Gómez-Salgado, J., Ortega-Moreno, M., y Navarro-Abal, Y. (2021). Relationship between work engagement, psychosocial risks, and mental health among spanish nurses: A cross-sectional Study. *Frontiers in Public Health*, 8, 627472.
- García-Rodríguez, A., Gutiérrez-Bedmar, M., Bellón-Saameno, J. Á., Muñoz-Bravo, C., y Navajas, J. F. C. (2015). Entorno psicosocial y estrés en trabajadores sanitarios de la sanidad pública: diferencias entre atención primaria y hospitalaria. *Atención Primaria*, 47(6), 359-366.
- Gómez-Sánchez, L. E. (2019). Séptima evaluación de test editados en España. *Papeles del Psicólogo*, 40(3), 205-210.
- Guillera, G., y Barrios, M. (2025). Undécima evaluación de test editados en España. *Papeles del Psicólogo*, 46(3), 158-166.
- Hernández, A., Elosua, P., Fernández-Hermida, J. R., y Muñiz, J. (2022). Comisión de Test: Veinticinco años velando por la calidad de los test. *Papeles del Psicólogo*, 43(1), 55-62.
- Hernández, A., Hidalgo, M. D., Hambleton, R. K., y Gómez Benito, J. (2020). International test commission guidelines for test adaptation: A criterion checklist. *Psicothema*, 32(3), 390-398.
- Hernández, A., Paradell, È., Saavedra, L., y Vallar, F. (2024a). *Test de vocabulario en imágenes Peabody-5: PPVT-5* [Adaptación española]. Pearson Educación.
- Hernández, A., Paradell, È., Saavedra, L., y Vallar, F. (2024b). *Test de vocabulario expresivo-3: EVT-3* [Adaptación española]. Pearson Educación.
- 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*, 37(1), 192-197.
- Hernández, A., Tomás, I., Ferreres, A., y Lloret, S. (2015). Tercera evaluación de tests editados en España. *Papeles del Psicólogo*, 36(1), 1-8.
- 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.
- International Labour Organization (2016). *Workplace Stress: A Collective Challenge*. En <https://www.ilo.org/publications/workplace-stress-collective-challenge>

- International Test Commission. (2018). ITC guidelines for translating and adapting tests. *International Journal of Testing*, 18(2), 101-134.
- International Test Commission y Association of Test Publishers (2022). *Guidelines for technology based assessment*. En <https://www.intestcom.org/page/28>
- Koo, T. K., y Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155-163.
- Kristensen, T. S., Hannerz, H., Høgh, A., y Borg, V. (2005). The Copenhagen Psychosocial Questionnaire-a tool for the assessment and improvement of the psychosocial work environment. *Scandinavian Journal of Work, Environment & Health*, 31(6), 438-449.
- Leka, S., Jain, A., y World Health Organization. (2010). *Health impact of psychosocial hazards at work: an overview*. World Health Organization.
- Louzán, R., y Villarroja López, A. (2024). Impact of social desirability bias on the CoPsoQ-ISTAS 21 psychosocial risk assessment questionnaire. *Work*, 79(1), 241-251.
- Lozano, L. M. (2023). Novena evaluación de los test editados en España. *Papeles del Psicólogo*, 44(1), 1-7.
- Moncada, S., Llorens, C., Font, A., Galtés, A., y Navarro, A. (2008). Exposición a riesgos psicosociales entre la población asalariada en España (2004-05): valores de referencia de las 21 dimensiones del cuestionario COPSOQ ISTAS21. *Revista Española de Salud Pública*, 82, 667-675.
- Moncada, S., Llorens, C., Moreno, N., y Molinero, E. (2014a). *Manual del método CoPsoQ-istas21 (versión 2) para la evaluación y la prevención de los riesgos psicosociales en empresas con 25 o más trabajadoras*. Instituto Sindical de Trabajo, Ambiente y Salud.
- Moncada, S., Llorens, C., Navarro, A., y Kristensen, T. S. (2005). ISTAS21: Versión en lengua castellana del cuestionario psicosocial de Copenhague (COPSOQ). *Archivos de Prevención de Riesgos Laborales*, 8(1), 18-29.
- Moncada, S., Llorens, C., Salas, S., Moriña, D., y Navarro, A. (2021). La tercera versión de COPSOQ-ISTAS21. Un instrumento internacional actualizado para la prevención de riesgos psicosociales en el trabajo. *Revista Española de Salud Pública*, 95, e1-16.
- Moncada, S., Utzet, M., Molinero, E., Llorens, C., Moreno, N., Galtés, A., y Navarro, A. (2014b). The Copenhagen Psychosocial Questionnaire II (COPSOQ II) in Spain-A tool for psychosocial risk assessment at the workplace. *American Journal of Industrial Medicine*, 57(1), 97-107.
- Muñiz, J. (2010). La Teoría de los Tests: Teoría Clásica y Teoría de Respuesta a los Ítems. *Papeles del Psicólogo*, 31(1), 57-66.
- Muñiz, J. (2018). La medición de lo psicológico. En Academia de Psicología de España (Ed.), *Psicología para un mundo sostenible. Volumen II* (pp. 13-29). Ed. Pirámide.
- Muñiz, J., Elosua, P., y Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los test: Segunda edición. *Psicothema*, 25(2), 151-157.
- Muñiz, J., y Fernández-Hermida, J. R. (2000). La utilización de los test en España. *Papeles del Psicólogo*, 76, 41-49.
- Muñiz, J., y Fernández-Hermida, J. R. (2010). La opinión de los psicólogos españoles sobre el uso de los test. *Papeles del Psicólogo*, 31(1), 108-121.
- Muñiz, J., Fernández-Hermida, J. R., Fonseca-Pedrero, E., Campillo-Álvarez, A., y Peña-Suárez, E. (2011). Evaluación de test editados en España. *Papeles del Psicólogo*, 32(2), 113-128.
- 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.
- Navarro, A., Fernández-Cano, M. I., Salas-Nicas, S., Llorens, C., Moriña, D., y Moncada, S. (2022). Relación entre exposición a riesgos psicosociales y salud: un estudio de cohorte mediante el COPSOQistas21. *Gaceta Sanitaria*, 36(4), 376-379.
- Ponsoda, V., y Hontangas, P. (2013). Segunda evaluación de test editados en España. *Papeles del Psicólogo*, 34(2), 82-90.
- Prieto, G., y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo*, 77, 65-77.
- Recio-Saboya, P. (2026). Duodécima evaluación de test editados en España. *Papeles del Psicólogo*, 47(2), 106-114.
- Reise, S. P., Bonifay, W. E., y Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, 95(2), 129-140.
- Sánchez-Sánchez, F., y Solar, M. (2024). *Cuestionario Para la Evaluación de las Funciones Ejecutivas y el TDAH: ATENTO*. Hogrefe TEA Ediciones.
- Schütte, S., Chastang, J. F., Malard, L., Parent-Thirion, A., Vermeylen, G., y Niedhammer, I. (2014). Psychosocial working conditions and psychological well-being among employees in 34 European countries. *International Archives of Occupational and Environmental Health*, 87, 897-907.
- Useche, S. A., Montoro, L., Alonso, F., y Pastor, J. C. (2019). Psychosocial work factors, job stress and strain at the wheel: Validation of the Copenhagen Psychosocial Questionnaire (COPSOQ) in professional drivers. *Frontiers in Psychology*, 10, 1531.
- Utzet, M., Moncada, S., Molinero, E., Llorens, C., Moreno, N., y Navarro, A. (2014). The changing patterns of psychosocial exposures at work in the South of Europe: Spain as a labor market laboratory. *American Journal of Industrial Medicine*, 57(9), 1032-1042.
- Viladrich, C., Doval, E., Penelo, E., Aliaga, J., Espelt, A., García-Rueda, R., y Angulo-Brunet, A. (2021). Octava evaluación de test editados en España: Una experiencia participativa. *Papeles del Psicólogo*, 42(1), 1-9.
- Williams, K. T. (2019). *Expressive Vocabulary Test: EVT-3* (3rd ed.). NCS Pearson.

Anexo. Listado de Nombres Completos de los Test Incluidos en la Tabla 1

Edición	Acronimo	Nombre
2024	ATENTO	Cuestionario para la Evaluación de las Funciones Ejecutivas y el TDAH
	BFQ-2	Big Five Questionnaire - 2
	CoPsoQ-Istas21-V2	Cuestionario Psicosocial de Copenhague Versión 2
	EBE	Batería de Evaluación del Bienestar Educativo
	EVT-3	Test de vocabulario expresivo-3
2023	PPVT-5	Test de vocabulario en imágenes Peabody-5
	APM	Análisis de Perfiles Motivacionales
	Matrices-TAI	Test Adaptativo de Inteligencia General (online)
	MACI-II	Inventario Clínico para adolescentes Millon-II
	SBE	Screening de bienestar emocional para niños y adolescentes
2022	SPMSQ	Cuestionario de Pfeiffer de evaluación de deterioro cognitivo
	TECO	Prueba para la evaluación del conocimiento fonológico
	THINEME	Test de Inteligencia Emocional
	BESS	Sistema de cribado conductual y emocional
	CTC-R	Cuestionario TEA Clínico-Revisado
2021	CTS-2	Versión revisada de la Conflict Tactics Scales
	CUMANIN-2	Cuestionario de madurez neuropsicológica infantil-2
	MASC2	Escala de Ansiedad Multidimensional para Niños/as
	SRP 4	Escala de Psicopatía
	TALE	Test de análisis de lectoescritura
2020	BASC-3.	Sistema de evaluación de la conducta de niños y adolescentes
	CIT	Cuestionario de Impacto del Trauma
	CPF	Cuestionario de Personalidad Forense
	CTE	Cuestionario de Talento Emprendedor
	DABS	Escala de Diagnóstico de Conducta Adaptativa
2019	GDS	Escala de Depresión Geriátrica
	PROLEXIA	Diagnóstico y Detección Temprana de la Dislexia
	BAYLEY-III	Escala Bayley de Desarrollo Infantil III
	BECOLE-R	Batería de Evaluación Cognitiva de las Dificultades en Lectura y Escritura. Revisada y Renovada
	CAG	Cuestionario de Autoconcepto Garley
2018	DAS	Escala de Ajuste Diádico
	MacArthur	MacArthur Inventario de Desarrollo Comunicativo
	RAVEN 2	Matrices Progresivas de Raven 2
	BADyG/i	Batería de Actividades mentales Diferenciales y Generales, Nivel infantil
	BADyG/E1-r	Batería de Actividades mentales Diferenciales y Generales, Nivel E1 renovado
2017	BADyG/E2-r	Batería de Actividades mentales Diferenciales y Generales, Nivel E2 renovado
	BADyG/E3-r	Batería de Actividades mentales Diferenciales y Generales, Nivel E3 renovado
	BADyG/M-r	Batería de Actividades mentales Diferenciales y Generales, Nivel M renovado
	BADyG/S-r	Batería de Actividades mentales Diferenciales y Generales, Nivel S renovado
	BRIEF-P	Evaluación Conductual de la Función Ejecutiva - Versión infantil
2016	CELF-5	Evaluación Clínica de los Fundamentos del Lenguaje, 5
	MCMI-IV	Inventario Clínico Multiaxial de Millon, IV
	PECO	Prueba para la Evaluación del Conocimiento Fonológico
	TONI-4	Test de Inteligencia No Verbal, 4
	BRIEF-2	Evaluación Conductual de la Función Ejecutiva -2
2015	BYI-2	Inventarios de Beck para niños y adolescentes - 2
	DP-3	Perfil de Desarrollo - 3
	Factor g-R	Test de Inteligencia No Verbal - Revisado
	IAES-A	Inventario de Ansiedad Escolar - Abreviado
	PAIB-1-R	Prueba de Aspectos Instrumentales Básicos en Lenguaje y Matemáticas: E. Infantil (5 años), 1º y 2º E. Primaria - Renovada
2014	PAIB-2-R	Prueba de Aspectos Instrumentales Básicos en Lenguaje y Matemáticas: 3º y 4º E. Primaria - Renovada
	PAIB-3-R	Prueba de Aspectos Instrumentales Básicos en Lenguaje y Matemáticas: 5º y 6º E. Primaria y 1º ESO - Renovada
	BADYG/E2-r	Batería de Aptitudes Diferenciales y Generales Renovado E2
	BADYG/S-r	Batería de Aptitudes Diferenciales y Generales
	BAT-7	Batería de Aptitudes de TEA
2013	BDI-FastScreen	Inventario de Depresión de Beck FastScreen para pacientes Médicos
	BPR	Batería de Pruebas de Razonamiento
	CESPRO	Batería para la Evaluación de las Estructuras Sintáctico-Semánticas que componen los enunciados de los problemas matemáticos y de la utilización de estrategias algorítmicas para su resolución
	MATRICES	MATRICES Test de Inteligencia General
	MBMD	Inventario Conductual de Millon para pacientes con diagnóstico médico
2012	Q-PAD	Cuestionario para la evaluación de problemas en adolescentes
	SENSORY PROFILE	Perfil Sensorial-2

Edición	Acrónimo	Nombre
2016	BADYG-E3	Batería de Aptitudes Diferenciales y Generales Renovado E3
	CAEPO	Cuestionario de Afrontamiento al Estrés para Pacientes Oncológicos
	EDI-3	Inventario de Trastornos de la Conducta Alimentaria-3
	EVAPROMES	Evaluación de los Procesos Metacognitivos en Escritura
	LAEA	Listado de Adjetivos para la Evaluación del Autoconcepto
	MABC-2	Batería de Evaluación del Movimiento para niños-2
	MABC-2-LOC	Lista de Observación Conductual de la MABC-2
	NEPSY-II	Batería Neuropsicológica Infantil-II
	PAIB 2 y 3	Prueba de Aspectos Instrumentales Básicos 2 y 3
	PRO 1-2 y 3	Protocolo de Evaluación del Lenguaje
	TEMT	Test de Evaluación Matemática Temprana
2015	WISC-V	Escala de Inteligencia de Wechsler para Niños-V
	ABAS-II	Sistema de Evaluación de la Conducta Adaptativa
	BADyG/M-r	Batería de Aptitudes Diferencial y Generales. Nivel M Renovado
	BETA	Batería para la Evaluación de los Trastornos Afásicos
	BSI-18	Inventario Breve de 18 Síntomas
	CECAD	Cuestionario Educativo-Clinico: Ansiedad y Depresión
	EHPAP	Evaluación de Habilidades y Potencial de Aprendizaje para Preescolares
	PAIB-1	Prueba de Aspectos Instrumentales Básicos en Lenguaje y Matemáticas
	PECC	Prueba para la Evaluación de la Cognición Cotidiana
	SCIP	Screening del Deterioro Cognitivo en Psiquiatría
	WMS-IV	Escala de Memoria de Wechsler -IV
2014	WPPSI-IV	Escala de Inteligencia de Wechsler para Preescolar y Primaria -IV
	BCSE	Test Breve para la evaluación del estado cognitivo
	BECOLE	Batería de Evaluación Cognitiva de la Lectura y la Escritura
	BOHEM-3	Test Boehm de conceptos básicos
	BOHEM-3 Preescolar	Test Boehm de conceptos básicos - 3 Preescolar
	CESQT	Cuestionario para la Evaluación del Síndrome de Quemarse por el Trabajo
	ECLE	Evaluación de la comprensión lectora
	ESQUIZO-Q	Cuestionario Oviedo para la Evaluación de la Esquizotipia
	IECI	Inventario de Estrés Cotidiano Infantil
	SOC	Escala de Dificultades de Socialización de Cantoblanco
	TRAUMA	Test de Resistencia al Trauma
2013	WAIS-IV	Escala de inteligencia de Wechsler para adultos-IV
	BAI	Inventario de ansiedad de Beck
	BAS-II	Escalas de aptitudes intelectuales
	BDI-II	Inventario de depresión de Beck-II
	CEAM	Cuestionario de estrategias de aprendizaje y motivación
	CompeTEA	Cuestionario para la evaluación de las competencias en el contexto laboral
	EPV-R	Escala de predicción del riesgo de violencia grave contra la pareja - Revisada
	ESCOLA	Escala de conciencia lectora
	ESPERI	Cuestionario para la detección de los trastornos del comportamiento en niños y adolescentes
	MerrillPalmer-R	Escalas de desarrollo Merrill-Palmer revisadas
	PAI	Inventario de evaluación de la personalidad
2012	RIAS / RIST	Escalas de inteligencia de Reynolds / Test de inteligencia breve de Reynolds
	WNV	Escala no verbal de aptitud intelectual de Wechsler
	16PF-5	Dieciséis Factores de Personalidad, quinta edición
	EFAI	Evaluación Factorial de la Aptitudes Intelectuales
	EVALUA	Batería Psicopedagógica
	IGF	Batería de Inteligencia General y Factorial
	MMPI-2RF	Inventario Multifásico de la Personalidad de Minnesota-2 Reestructurado
	NEO PI-R	Inventario de Personalidad NEO Revisado
	PROLEC-R	Batería de Evaluación de procesos Lectores, revisada
	WISC-IV	Escala de inteligencia de Wechsler para niños-IV