

Artículo

Undécima Evaluación de Test Editados en España

Georgina Guillera  y Maite Barrios 

Universidad de Barcelona, España

INFORMACIÓN

Recibido: Enero 22, 2025

Aceptado: Abril 30, 2025

Palabras clave

Test
Evaluación de test
Psicometría
Propiedades psicométricas
CET-R

RESUMEN

Los test psicológicos son herramientas esenciales en la psicología aplicada, utilizadas en diversos contextos para respaldar decisiones. En España, la Comisión de Test del Consejo General de la Psicología lidera la evaluación de pruebas desde 2010 mediante el Cuestionario para la Evaluación de Test-Revisado (CET-R). Este trabajo presenta los resultados de la undécima edición (2022-2024), en la que se han evaluado siete test: seis de diferentes casas editoriales y un test no comercial. Además, se compara el reporte de información y calidad psicométrica de los test editados antes y después del inicio de la implementación de este sistema de evaluación. Los resultados de la undécima evaluación muestran que los test comerciales y publicados o actualizados más recientemente presentan, en general, mejores valoraciones. Respecto a la comparación en la información reportada, los resultados muestran un aumento en el reporte de propiedades como el funcionamiento diferencial de los ítems y la estabilidad temporal, aunque no se observaron mejoras estadísticamente significativas en las puntuaciones de calidad. En general, el sistema fomenta la mejora continua de los test, promoviendo su actualización técnica y fortaleciendo la transparencia y confianza en los instrumentos utilizados por los psicólogos.

Eleventh Assessment of Test Edited in Spain

ABSTRACT

Psychological tests are essential tools in applied psychology, widely used in various contexts to support decisions. In Spain, the Test Commission of the General Council of Psychology has led the evaluation of tests since 2010 using the Test Evaluation Questionnaire-Revised (CET-R). This study presents the results of the eleventh edition (2022-2024), in which seven tests were evaluated: six from different publishing houses and one non-commercial test. Additionally, it compares the reporting of information and psychometric quality of tests published before and after the implementation of this evaluation system. The results of the eleventh evaluation show that commercially available tests, especially those that have been published or updated recently, generally receive higher ratings. Regarding the reporting of information, findings indicate an increase in the inclusion of properties such as differential item functioning and temporal stability, although no statistically significant improvements were observed in quality scores. Overall, the system fosters the continuous improvement of tests by promoting technical updates and strengthening transparency and confidence in the instruments used by psychologists.

Keywords

Tests
Test evaluation
Psychometrics
Psychometric properties
CET-R

Introducción

Los test psicológicos son herramientas fundamentales en el ejercicio profesional de la psicología, utilizadas de manera frecuente para respaldar decisiones en diversos campos como el clínico, educativo, organizacional y de la salud. Estas pruebas permiten diagnosticar, seleccionar, orientar y adaptar intervenciones, y su valor radica en su naturaleza estandarizada y en la solidez de sus propiedades psicométricas, que proporcionan precisión y confianza en los resultados (Muñiz et al., 2020). Sin embargo, debido al gran número de test disponibles y a las importantes consecuencias que pueden derivarse de sus aplicaciones, es esencial que los psicólogos tengan acceso a información contrastada y detallada sobre la calidad de estos instrumentos para poder elegir los más adecuados en cada contexto de evaluación.

En respuesta a esta necesidad, organismos internacionales, como la *American Psychological Association* (APA; Asociación Americana de Psicología), junto con la *American Educational Research Association* (AERA, Asociación Americana de Investigación Educativa) y el *National Council on Measurement in Education* (NCME, Consejo Nacional de Medición en Educación), han desarrollado estándares rigurosos para garantizar la calidad y el uso adecuado de los test (AERA, APA, y NCME, 2014). Otros, como la *International Test Commission* (ITC, Comisión Internacional de Test), han elaborado directrices sobre la traducción y adaptación de test (Hernández et al., 2020; ITC, 2017; Muñiz et al., 2013) y sobre el uso de test en general (ITC, 2013) o en ámbitos concretos (e.g., investigación: ITC, 2014; entornos digitales: ITC y Association of Test Publishers, 2022). En cuanto a la evaluación de la calidad de los test, en Europa, la *European Federation of Psychologists' Associations* (EFPA; Federación Europea de Asociaciones de Psicólogos) ha desarrollado e implementado un modelo de evaluación exhaustivo para valorar la validez y fiabilidad de las inferencias los test utilizados en la región (Evers et al., 2013), el cual se encuentra en proceso de actualización (Schittekatte y Evans, 2023). En el contexto español, la Comisión de Test del Consejo General de la Psicología (COP) ha asumido el liderazgo en la evaluación sistemática de los test editados en España desde el año 2010 (Hernández et al., 2022). Esta iniciativa surge como respuesta a la demanda de los psicólogos por disponer de evaluaciones independientes que les ofrezcan información técnica y práctica sobre la calidad de los instrumentos que emplean en su práctica diaria (Muñiz et al., 2011; Hernández et al., 2016). Para llevar a cabo estas evaluaciones, se ha empleado el Cuestionario de Evaluación de Test (CET), el cual fue desarrollado inicialmente por Prieto y Muñiz (2000) y ha sido modificado en diversas ocasiones para alinearse con los modelos europeos e incorporar los avances tecnológicos y prácticas de análisis psicométricos más actuales (Hernández et al., 2016).

Hasta la fecha, se han llevado a cabo múltiples ediciones de la evaluación de test en España, cada una de las cuales ha resultado en la publicación de informes detallados accesibles a los profesionales a través de la página web del COP (<https://www.cop.es/test/>). La importancia de estas evaluaciones es evidente en los resultados de estudios como los de Muñiz et al. (2020), que muestran un uso creciente de los test por parte de los psicólogos españoles, quienes valoran positivamente la utilidad de estos para la toma de decisiones en contextos diversos. Sin embargo, también subrayan la necesidad de contar con más información técnica que

les permita tomar decisiones basadas en la evidencia sobre la adecuación de los test para usos concretos. En este sentido, únicamente el 22,5% de los psicólogos colegiados encuestados en el estudio de Muñiz y colaboradores afirmó conocer la evaluación anual realizada por la Comisión de Test, aunque aquellos que sí eran conocedores consideraron que estas evaluaciones son útiles y necesarias, y afirmaron que consultan los informes para decidir qué test utilizar en su práctica profesional (Muñiz et al., 2020). Más allá de los beneficios que este proceso de evaluación puede suponer para los psicólogos que aplican los test, sería deseable que también contribuyera a mejorar la calidad de las pruebas en sí mismas, promoviendo un desarrollo continuo en su construcción, actualización y aplicación.

El presente manuscrito expone, por un lado, los resultados de la undécima edición del proceso de evaluación de test editados en España, llevado a cabo a lo largo de los años 2022-2024. Por otro lado, se presentan los resultados del análisis comparativo del reporte de información y la calidad psicométrica de los test antes y después del inicio de la implementación del sistema de evaluación.

Undécima Evaluación de Test

En este apartado se presentan los test sometidos a evaluación, el proceso de evaluación seguido, las personas que participaron como revisores y los principales resultados de la undécima evaluación de test editados en España.

Test Sometidos a Evaluación

Siguiendo el procedimiento que se ha adoptado en ediciones anteriores, las editoriales con representación en la Comisión de Test (i.e., Editorial CEPE, Giunti Psychometrics, Hogrefe TEA Ediciones y Pearson Clinical Assessment España) propusieron a la comisión los seis test que querían someter a evaluación. Adicionalmente, la Comisión de Test acordó incluir un séptimo test sin comercializar en España, la versión revisada de la *Conflict Tactics Scales* (CTS-2; Straus et al., 1996), adaptada al español por Loinaz (2009) y validada en el trabajo de Loinaz y colegas (Loinaz et al., 2012).

De acuerdo con esta forma de proceder, en esta undécima edición se han evaluado siete pruebas que abarcan diferentes áreas de la psicología aplicada, tales como la psicología clínica, educativa, forense y del trabajo y las organizaciones, pero también otros ámbitos como la neuropsicología o la investigación. En la *Tabla 1* se listan los siete instrumentos sometidos a revisión, seis de los cuales se encuentran comercializados por editoriales.

Selección de Revisores

Basándose en las ediciones anteriores de evaluación de test, con la intención de incorporar para cada test un revisor experto en aspectos técnicos y psicométricos y otro con experiencia en las variables evaluadas por los test, se buscaron potenciales revisores combinando varias estrategias de búsqueda. Para la identificación de expertos en psicometría, (a) se consultaron las listas de revisores de las diez ediciones anteriores de revisión de test publicadas en *Papeles del Psicólogo* (Abad, 2024; Elosua y Geisinger, 2016; Fonseca y Muñiz, 2017; Gómez-Sánchez, 2019; Hernández et al., 2015; Hidalgo y Hernández, 2019; Lozano,

Tabla 1

Listado de los Instrumentos de Medida de la Undécima Edición de Test

Acrónimo	Test	Editorial	Autor/es	Año publicación/actualización
BESS	Sistema de cribado conductual y emocional del BASC-3	Pearson Educación	Departamento de I+D de Pearson Clinical & Talent Assessment, A. Hernández, È. Paradell y F. Vallar	2022
CTC-R	Cuestionario TEA Clínico - Revisado	Hogrefe TEA Ediciones	D. Arribas, S. Corral y J. Pereña	2022
CTS-2	Versión revisada de la Conflict Tactics Scales	--	I. Loinaz, E. Echeburúa, M. Ortiz-Tallo y P. J. Amor	2012
CUMANIN-2	Cuestionario de madurez neuropsicológica infantil-2	Hogrefe TEA Ediciones	J. A. Portellano, R. Mateos, R. Martínez-Arias, F. Sánchez-Sánchez	2021
MASC2	Escala de Ansiedad Multidimensional para Niños/as	Giunti Psychometrics España	Equipo R&D - Giunti Psychometrics España, A. Martínez, J. Miralles e I. de Ancos	2022
SRP 4	Escala de Psicopatía	Giunti Psychometrics España	Equipo R&D - Giunti Psychometrics España, A. Martínez, J. Miralles e I. de Ancos	2021
T.A.L.E.	Test de análisis de lectoescritura	Machado Grupo de Distribución, S.L.	J. Toro y M. Cervera	1984

2023; Muñiz et al., 2011; Ponsoda y Hontangas, 2013; Viladrich et al., 2021); (b) se extrajeron los datos de contacto de los miembros de la *European Association of Methodology* (EAM, Asociación Europea de Metodología) con afiliación española (fecha de descarga desde la web de la asociación: 30 de junio de 2022); y, finalmente, (c) se completó la lista con otros investigadores expertos en psicometría y metodología, identificados dentro de la red de contactos de las coordinadoras. En el caso de los expertos en el constructo, para su identificación, (a) se consultaron también las listas de los revisores de ediciones anteriores; y (b) se realizaron búsquedas en bases de datos electrónicas utilizando, por un lado, el nombre del test y, por otro, el constructo que medía cada uno de los test. Inicialmente, se contactó con 14 revisores (i.e., 7 psicómetras y 7 expertos en el constructo), de los cuales un 28,6% (i.e., 4 psicómetras) aceptó en primera instancia; las vacantes fueron reemplazadas por nuevos revisores. En la segunda llamada a colaborar, destinada a cubrir las vacantes con nuevos revisores, se logró completar el perfil de psicómetras para todos los test. Sin embargo, fue necesario realizar múltiples intentos para garantizar la participación de expertos en el constructo dentro del proceso de revisión. En concreto, mientras que para el perfil de psicómetras se contactó con 11 investigadores, para cubrir el perfil de expertos en el constructo fue necesario contactar con 17. En la Tabla 2 se presenta la lista de revisores que participó en la undécima evaluación de test.

Instrumento de Evaluación: el CET-R V1.1

Para la evaluación de los test, se empleó el Cuestionario para la Evaluación de Test Revisado (CET-R; Hernández et al., 2016), diseñado específicamente para describir y valorar la calidad técnica y psicométrica de las pruebas, en su versión más actual (i.e., versión V1.1, disponible en la página web del COP: <https://www.cop.es/test/>). En el trabajo de Abad (2024) se describe detalladamente la estructura en secciones del instrumento (i.e., Descripción general del test, Valoración de las características del test y Valoración global del test), los aspectos del test evaluados en cada sección —ya sea mediante preguntas abiertas o cerradas—, así como el sistema de

Tabla 2

Listado de Revisores de la Undécima Evaluación de Test

Nombre y apellidos	Filiación
Jone Aliri Lazcano	Universidad del País Vasco - Euskal Herriko Unibertsitatea (UPV - EHU)
Isabel Benítez Baena	Universidad de Granada (UGR)
Carlos García Forero	Universitat Internacional de Catalunya (UIC)
Ana Martina Greco	Universitat Oberta de Catalunya (UOC)
Francisco Pablo Holgado Tello	Universidad Nacional de Educación a Distancia (UNED)
María Ángeles Jurado Luque	Universitat de Barcelona (UB)
Beatriz Molinuevo Alonso	Universitat Autònoma de Barcelona (UAB)
Mireia Orgilés Amorós	Universitat Miguel Hernández (UMH)
Eva Penelo Werner	Universitat Autònoma de Barcelona (UAB)
Jordi Renom Pinsach	Universitat de Barcelona (UB)
Susana Sanduvete	Universidad de Sevilla (US)
Paz Suárez Coalla	Universidad de Oviedo (UniOvi)
Jorge Torres Marín	Universidad de Granada (UGR)
Rafael Torrubia Beltri	Universitat Autònoma de Barcelona (UAB)

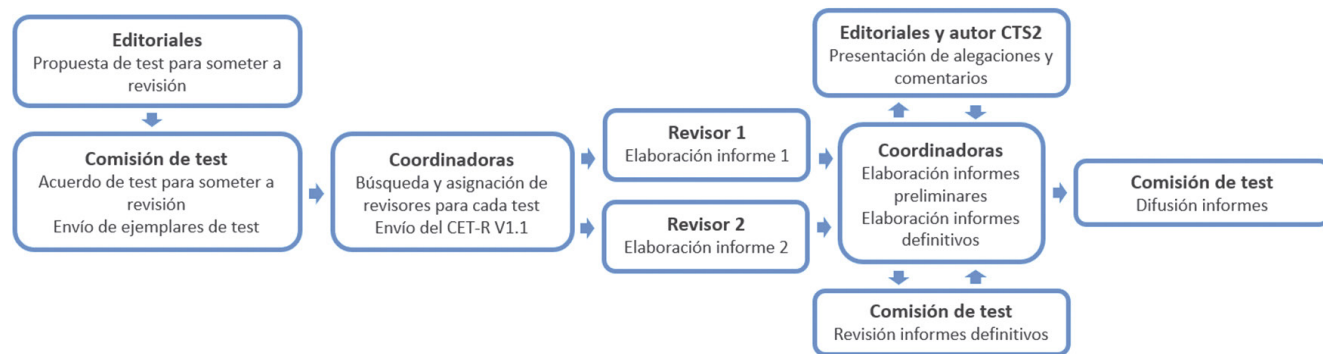
puntuación y las etiquetas correspondientes de los ítems cuantitativos. En esta última versión del instrumento, que ya se utilizó en Abad (2024), se introdujeron mejoras significativas, como (1) una distinción más clara entre la información fundamental para evaluar la calidad de un test y aquella que, aunque ausente, no resulta fundamental dada la finalidad del test; (2) criterios más específicos para pruebas adaptadas de otros idiomas y/o culturas, en los que se solicita a los revisores que indiquen la procedencia de las muestras (i.e., locales, internacionales o mixtas) en los diferentes apartados y que otorguen mayor peso a las evidencias obtenidas con muestras locales; y (3) orientaciones para valorar el Área Bajo la Curva (AUC) en el uso de curvas ROC, un aspecto clave en la predicción de criterios diagnósticos.

Proceso de Evaluación

Por lo que concierne a las seis pruebas comercializadas, una vez que los revisores aceptaron participar, el personal administrativo

Figura 1

Procedimiento de Revisión de Test Seguido en la Undécima Edición



del COP se encargó de gestionar el envío de las pruebas a cada uno de los dos revisores, así como a las responsables de la coordinación del proceso de evaluación. Paralelamente, la coordinadora de la Comisión de Test envió a las encargadas de la presente edición de evaluación de test los datos de contacto de los representantes de las editoriales y el CET-R V1.1. En el caso de la versión española de la CTS-2, al no encontrarse comercializada, se realizó una búsqueda y localización de trabajos combinando diversas estrategias. Primero, se buscaron trabajos que citaban el estudio de validación de la versión española (Loinaz et al., 2012) en *Scopus* ($n = 43$) y *Web of Science* ($n = 45$) [búsqueda realizada en julio de 2022]. Segundo, se buscaron trabajos que en el título, el resumen o las palabras clave contenían el nombre de la escala “Conflict Tactics Scales-2” en *Scopus* ($n = 63$) y *Web of Science* ($n = 18$) [búsqueda realizada en julio de 2022]. Tras la lectura del título y el resumen y, en caso necesario, del texto completo, se seleccionaron aquellos trabajos que, a juicio de las coordinadoras, aportaban información relevante para valorar la calidad del test (e.g., estudios empíricos y de revisión sobre propiedades psicométricas de la escala). Tercero, revisando las referencias de éstos, se identificaron siete trabajos adicionales (i.e., Corral y Calvete, 2006; Espejo-Navarro y Valdivia-Ramírez, 2023; Gallego Rodríguez y Fernández-González, 2019; Graña et al., 2013; Loinaz, 2009; Muñoz-Sánchez, 2018; Redondo y Graña, 2015). Este proceso de búsqueda y localización de trabajos terminó con la creación de una lista de 50 documentos, mayoritariamente artículos que emplearon la versión española de la CTS-2, aunque también incluyó artículos sobre la versión original en inglés (e.g., Straus et al., 1996), así como revisiones de las propiedades psicométricas de la escala (e.g., Chapman y Gillespie, 2019). Esta documentación se facilitó a los dos revisores de la CTS-2.

Tras la recepción de la documentación, los revisores procedieron a la evaluación del test, de forma independiente, a través del CET-R V1.1. Los informes por parte de los revisores se recibieron hasta abril de 2023. Para cada test, las coordinadoras integraron las evaluaciones de los dos revisores asignados y elaboraron informes preliminares, los cuales fueron enviados a las editoriales y el autor de correspondencia del estudio de validación de la versión española de la CTS-2 para brindarles la oportunidad de presentar alegaciones. Finalmente, las coordinadoras prepararon los informes definitivos, los cuales fueron revisados por un miembro de la Comisión de Test. Sus sugerencias, principalmente de estilo y aclarativas, fueron incorporadas para la publicación de la versión final de los informes

en la web del COP en mayo de 2024. En la Figura 1 se muestra un esquema del procedimiento de revisión de test seguido en la undécima edición.

Resultados

En la Tabla 3 se muestran las puntuaciones resultantes del proceso de evaluación otorgadas a los distintos test junto con la puntuación media de todas las características valoradas. En general, todas las puntuaciones medias reflejan valoraciones entre buenas y excelentes para los test comerciales publicados o actualizados más recientemente. En cambio, el CTS-2 —test no comercial— y el T.A.L.E. —el cual no ha sido actualizado en las últimas cuatro décadas— presentan puntuaciones inadecuadas o, en el mejor de los casos, adecuadas, pero con carencias.

Análisis Comparativo del Reporte de Información y la Calidad de los Test Antes y Después del Inicio del Proceso de Evaluación

Objetivo

Tal como se ha señalado, el proceso de revisión de test editados en España, liderado por la Comisión de Test del COP, nació para dar respuesta a una de las principales demandas de los psicólogos profesionales: disponer de información técnica que facilite la toma de decisiones fundamentadas. Tras completar once procesos de evaluación, planteamos la posibilidad de analizar comparativamente el reporte de información y la calidad de los test publicados antes y después del inicio del sistema de evaluación, atendiendo a los resultados de haber aplicado el CET, en sus distintas versiones, a lo largo de las once ediciones. Se parte del supuesto de que la introducción del sistema de evaluación debería haber repercutido a dos niveles. Por un lado, a que se reporte más información de las propiedades psicométricas de los test y, por otro, a que los test mejoren las valoraciones que reciben, dado que se espera que los creadores de test tengan en consideración los estándares y recomendaciones propuestos por la Comisión de Test, aunque sea de forma paulatina. Concretamente, se explora hasta qué punto la implementación del sistema de evaluación ha supuesto un aumento de (1) la información proporcionada en cuanto a las características psicométricas de los test (cantidad de información), y (2) las puntuaciones en cada una de las características de los test evaluadas (calidad de los test).

Tabla 3*Puntuaciones de los Test Evaluados en la Undécima Edición*

Características	BESS	CTC-R	CTS-2	CUMANIN-2	MASC2	SRP4	T.A.L.E.
Desarrollo							
Materiales y documentación	4,5	5	-	5	4,5	3,8	4
Fundamentación teórica	4	5	3	4	4	5	1
Adaptación	4	-	3	-	3	3	-
Análisis de ítems	-	4	4	5	4	-	2
Validez							
Contenido	4	4	2,5	4,5	3	3,5	1,5
Relación con otras variables	3,9	4,4	2,6	4,2	4	3,8	-
Estructura interna	2	4,5	2	5	4	3	-
Análisis del DIF	-	5	-	5	5	-	-
Fiabilidad							
Equivalencia	-	-	-	-	-	-	-
Consistencia interna	3,3	5	3	5	4	3,8	-
Estabilidad	3,5	3,5	2	4,3	3,5	2,3	-
TRI	-	-	-	3	4	-	-
Inter-jueces	-	-	-	-	-	-	-
Baremos e interpretación de las puntuaciones	4	5	3	5	3,5	3	1,7
Total	3,7	4,5	2,8	4,5	3,9	3,5	2,0

Nota. DIF: differential item functioning, funcionamiento diferencial del ítem; TRI: teoría de respuesta al ítem. Las puntuaciones corresponden a una escala valorativa de cinco puntos, donde: 1 = Inadecuada, 2 = Adecuada con algunas carencias, 3 = Adecuada, 4 = Buena y 5 = Excelente. El símbolo (-) indica que no se aporta información o no procede por las características del test.

Procedimiento

Se extrajeron las puntuaciones de los 96 informes disponibles en la página web de la Comisión de Test de las diez ediciones completadas hasta el momento, a las que se añadieron los correspondientes a los siete test de la presente edición, resultando en un total de 103 informes. Los tres test no comercializados se excluyeron del análisis: la Escala de Predicción del riesgo de Violencia grave contra la pareja-Revisada (EPV-R; [Echeburúa et al., 2010](#)), la *Geriatric Depression Scale - short version* (GDS; [Martínez de la Iglesia et al., 2002, 2005](#)) y la *Revised Conflict Tactics Scales* (CTS-2; [Loinaz, 2009](#); [Loinaz et al., 2012](#)).

Los test se clasificaron en tres categorías según su año de publicación o la fecha más reciente de actualización: (1) periodo anterior al inicio del proceso de evaluación de test por parte de la Comisión de Test, en el que se incluyeron test publicados hasta el 2011, año en el que se publicaron los resultados de la primera edición (i.e., [Muñiz et al., 2011](#)); (2) periodo de adaptación, definido como dos años de intervalo de tiempo durante el cual los desarrolladores de test tuvieron la oportunidad de familiarizarse y ajustarse a los estándares y criterios señalados en el modelo de evaluación, y que incluyó test publicados en los años 2012 y 2013; y (3) periodo posterior a la introducción del proceso de evaluación de test, el cual incluyó test publicados a partir del año 2014. De esta forma se dispuso de datos sobre test publicados con anterioridad ($n = 39$) y con posterioridad ($n = 50$) al inicio del proceso de evaluación de test, cuyas valoraciones fueron las que se incluyeron en el estudio comparativo (ver [Figura 2](#)).

De cada informe, se extrajo la información de la valoración general de los respectivos test (apartado Valoración general del CET V1.1), la cual incluye 14 características (ver [Tabla 3](#), columna Características). Se codificó, por un lado, si se reportaba la característica o propiedad psicométrica en cuestión (Reportada vs. No reportada) y, por otro, la puntuación específica para dicha característica o propiedad (valor numérico del 0 al 5). Dado que el

modelo del CET empleado a lo largo de las sucesivas ediciones ha experimentado algunas modificaciones, se utilizó como modelo de base las 14 características de la versión más actual del CET-R (i.e., CET-R V1.1). En la mayoría de ellas existe una correspondencia directa entre las versiones utilizadas en los diferentes informes (por ejemplo, en el caso de la evidencia basada en el contenido); sin embargo, en algunas se tuvo que establecer una correspondencia entre informes (por ejemplo, “Resultados del análisis factorial” se ubicó en “Validez: estructura interna” o “Validez predictiva” en “Validez: relaciones con otras variables”), lo que en algunos casos supuso la necesidad de calcular puntuaciones promedio.

En la comparación sobre la información reportada (i.e., Reportada vs. No reportada) entre el periodo anterior y posterior al inicio del sistema de evaluación, se aplicó la prueba χ^2 , la cual se acompañó del coeficiente ϕ como medida tamaño del efecto. Por lo que concierne al estudio comparativo de la calidad de los test, se aplicó la prueba U de Mann-Whitney con las puntuaciones recibidas en cada característica o propiedad psicométrica y se calculó el tamaño del efecto η^2 . La base de datos y la sintaxis empleadas en este análisis puede consultarse en el siguiente enlace: <https://osf.io/t43gf/>.

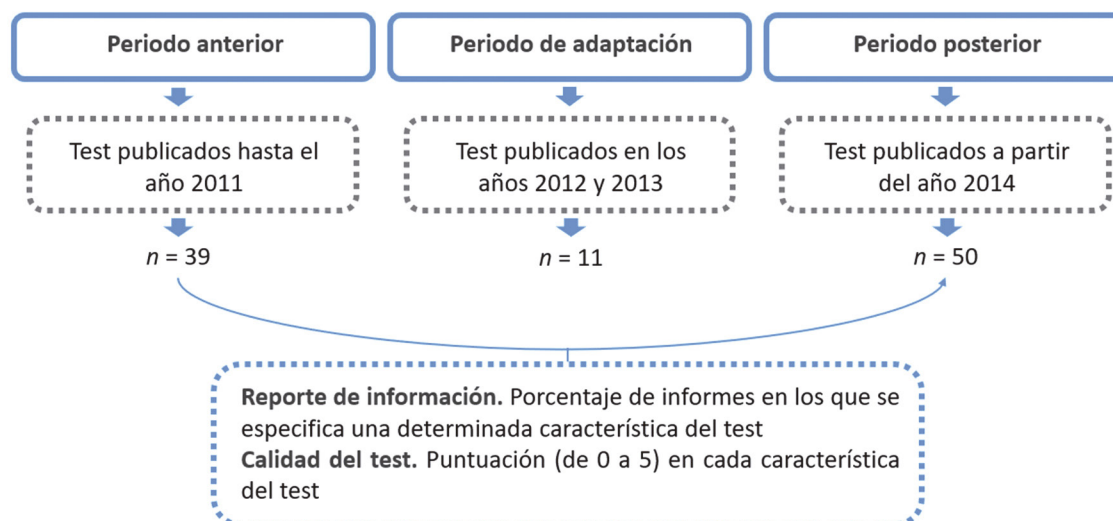
Resultados

Reporte de información. Tal y como se muestra en la [Tabla 4](#), se observó que, independientemente de la calificación obtenida, los test publicados con posterioridad a la introducción del proceso de revisión presentan, en un porcentaje superior, información respecto a algunas propiedades psicométricas (p.ej., el funcionamiento diferencial de los ítems o la estabilidad temporal).

Calidad de los test. En la [Tabla 4](#) se presentan las puntuaciones promedio obtenidas antes y después de implantar el sistema de evaluación, junto con el valor del estadístico de contraste y su correspondiente tamaño del efecto. Aunque para varias de las características analizadas el rango promedio de las puntuaciones

Figura 2

Esquema del Procedimiento Utilizado en la Comparación Entre el Periodo Anterior y Posterior al Inicio del Sistema de Evaluación de Test



fue superior en los test publicados con posterioridad al inicio del proceso de evaluación, la prueba U de Mann Whitney no resultó estadísticamente significativa en ningún caso (ver [Tabla 4](#)).

Si bien este análisis comparativo puede incentivar la reflexión sobre el proceso de revisión de test, no está exento de limitaciones. Uno de los principales desafíos durante la extracción de la información de los informes de cada test fue la necesidad de realizar algunos ajustes debido a los cambios realizados en las diferentes versiones del CET y, en consecuencia, en cómo queda registrada la información en los informes. Estos cambios obligaron a realizar modificaciones para extraer la información y las

puntuaciones correspondientes a las 14 características de la Valoración General. Por ejemplo, en los primeros informes correspondientes a la primera y segunda evaluación de test, no se pudo extraer una puntuación para 'Validez: Estructura interna'. De forma similar, en los informes relativos a la tercera y cuarta ronda de revisiones, el apartado 'Validez de constructo' incluía el análisis de la estructura interna, la comparación de grupos, la evidencia convergente y discriminante, entre otros aspectos. Sin embargo, en las versiones más recientes del CET, este apartado se dividió en 'Validez: Estructura interna' y 'Validez: Relaciones con otras variables', lo que obligó a asignar a esta característica la

Tabla 4

Comparación del Reporte de Información y Calidad de los Test Entre el Periodo Anterior ($n = 39$) y Posterior ($n = 50$) al Inicio del Sistema de Evaluación de Test

Características	Reporte (% Si)					Calidad (Media)				
	Anterior	Posterior	χ^2 (g.l.)	p valor	ϕ	Anterior	Posterior	Z_U	p valor	η^2
Desarrollo										
Materiales y documentación	100	100	-	-	-	4,2	4,4	-1,364	,172	0,02
Fundamentación teórica	100	100	-	-	-	4,1	4,3	-1,005	,315	0,01
Adaptación	100	100	-	-	-	4,3	4,3	-0,219	,827	< 0,01
Análisis de ítems	79,5	84,0	0,303(1)	,582	0,06	3,9	3,8	-0,582	,561	< 0,01
Validez										
Contenido	94,9	94,0	0,031(1)	,895	-0,02	3,8	3,9	-0,139	,890	< 0,01
Relación con otras variables	97,4	100	1,297(1)	,255	0,12	3,7	3,7	-0,173	,862	< 0,01
Estructura interna	84,6	84,0	0,006(1)	,937	-0,01	3,8	3,7	-0,019	,985	< 0,01
Análisis del DIF	10,3	30,0	5,087(1)	,024	0,24	4,4	4,1	0,000	1,000	0,00
Fiabilidad										
Equivalencia	100	100	-	-	-	3,7	3,5	-0,236	,814	0,01
Consistencia interna	97,4	100	1,297(1)	,255	0,12	4,1	4,4	-1,892	,059	0,04
Estabilidad	38,5	64,0	5,734(1)	,017	0,25	3,4	3,6	-0,469	,639	< 0,01
TRI	0,00	22,2	3,234(1)	,072	0,24	-	3,9	-	-	-
Inter-jueces	-	100	-	-	-	-	4,3	-	-	-
Baremos e interpretación de las puntuaciones	94,9	100	2,623	,105	0,17	3,9	4,1	-1,321	,186	0,02

Nota. DIF: differential item functioning, funcionamiento diferencial del ítem; TRI: teoría de respuesta al ítem. Las puntuaciones medias de calidad corresponden a una escala valorativa de cinco puntos, dónde: 1 = Inadecuada, 2 = Adecuada con algunas carencias, 3 = Adecuada, 4 = Buena y 5 = Excelente. El símbolo (-) indica que no se aporta información o no procede por las características del test.

puntuación de ‘Resultados del análisis factorial’. Además, algunas características y análisis psicométricos, como la ‘Fiabilidad: TRI’ y la ‘Fiabilidad inter-jueces’, se incorporaron al CET en etapas posteriores. Aunque es posible que la descripción y valoración de estos aspectos se incluyeran en las secciones abiertas de los informes, el presente análisis comparativo no incorporó la información contenida en dichas secciones. Por último, el periodo de adaptación se estableció en dos años. Sin embargo, este tiempo podría ser insuficiente para que los desarrolladores de test se familiarizaran y ajustaran a las nuevas exigencias. Estas limitaciones resaltan la necesidad de considerar posibles sesgos o carencias al interpretar los resultados obtenidos en el presente trabajo.

Conclusiones

Procedimiento de Evaluación

El proceso de evaluación de los test, liderado por la Comisión de Test del COP, sigue un enfoque riguroso y sistemático que garantiza la evaluación integral de las propiedades psicométricas y técnicas de los instrumentos de medición (Fernández-Ballesteros et al., 2001). Este procedimiento ha facilitado la integración, por un lado, de la opinión de revisores expertos en psicometría y en las variables evaluadas por los test y, por otro, de las recomendaciones y alegaciones de editoriales y autores, lo que refuerza la calidad de los informes publicados.

Cabe destacar que, aunque se logró completar los equipos de revisión para todos los test, el proceso evidenció diferencias en la facilidad de reclutar psicómetras frente a expertos en el constructo. Mientras que el perfil de psicómetras pudo ser cubierto de manera relativamente rápida, el reclutamiento de expertos en el constructo requirió un mayor esfuerzo y múltiples intentos, lo que podría reflejar una menor percepción de la importancia del proceso y la relevancia de los resultados de la revisión en su práctica profesional. Este hallazgo subraya la necesidad de implementar estrategias específicas para atraer y comprometer a estos especialistas, dada su relevancia en la evaluación integral de los instrumentos. Aunque el acuerdo entre revisores —psicómetras y expertos en contenido— de un mismo test varía según la característica del test evaluada (Abad, 2024), ambos perfiles contribuyen a que la valoración sea precisa y comprehensiva. Finalmente, resulta fundamental intensificar los esfuerzos de difusión del proceso de revisión de test basado en el CET en los distintos ámbitos aplicados de la psicología (Muñiz et al., 2020).

Los resultados obtenidos en esta undécima edición evidencian que los test comerciales de publicación o actualización más reciente alcanzan valoraciones medias que oscilan entre buenas y excelentes. En contraste, el CTS-2 —test no comercial— y el T.A.L.E. —sin revisiones en los últimos cuarenta años— registran puntuaciones que indican valoraciones inadecuadas o, como mucho, adecuadas, pero con carencias. Por un lado, conviene tener en cuenta que la revisión de pruebas no comerciales plantea retos importantes, como la falta de manuales y materiales, así como la dispersión de información relevante (Abad, 2024). Además, la heterogeneidad de los estudios en términos de diseño, tamaño y composición muestral, así como en cuanto a su calidad metodológica, podría estar contribuyendo a la variabilidad

observada en las valoraciones de las propiedades psicométricas de estas pruebas. En relación con el T.A.L.E., y considerando tanto su falta de actualización como los resultados de calidad obtenidos, se recomienda utilizar instrumentos más recientes para la evaluación de la lectoescritura. Un ejemplo son las Baterías de Evaluación Cognitiva de las Dificultades en la Lectura y Escritura (BECOLE-R; Galve Manzano y Martínez Arias, 2019), la cual fue evaluada en la novena edición del sistema de evaluación de test (Lozano, 2023), obteniendo valoraciones medias entre buenas y excelentes.

Análisis Comparativo del Reporte de Información y la Calidad de los Test

En la comparación entre el periodo anterior y posterior al inicio del proceso de evaluación de test, se evidencia un incremento significativo en la inclusión de algunas características psicométricas en los manuales, concretamente por lo que respecta al DIF y a la estabilidad temporal. Sin embargo, las características evaluadas no muestran mejoras estadísticamente significativas en sus puntuaciones de calidad.

Atribuir estos progresos exclusivamente al proceso de evaluación de test sería una conclusión, como mínimo, arriesgada, ya que son múltiples y variadas las causas que podrían explicar esta tendencia. Por ejemplo, el incremento de los estudios de DIF o del uso de la TRI podría reflejar lo que también ha sucedido a nivel internacional. Entre los posibles factores, destacan los avances informáticos y la mayor disponibilidad de software fácil de usar, que elimina la necesidad de programación avanzada, así como una creciente sensibilización hacia la perspectiva de género en la investigación, lo que podría haber impulsado más estudios de invarianza por género. Además, es probable que los desarrolladores de test en España estén incorporando de manera más habitual los estándares del CET, o que las editoriales utilicen el CET como modelo y aspiren a alcanzar el máximo nivel posible, viendo en ello un estándar de calidad ideal. Este enfoque podría estar motivado por el deseo de obtener buenas valoraciones en los informes y de presentar sus test como productos de alta calidad.

Sea como sea, no es descartable que este sistema de evaluación mediante el CET haya incentivado la actualización continua de los test, alentando a editoriales y autores a implementar análisis más avanzados y a fortalecer la fundamentación técnica de sus instrumentos. Estos avances no solo incrementan la transparencia y la confianza en la calidad de los test disponibles en España, sino que también facilitan la toma de decisiones informadas por parte de los profesionales de la psicología, contribuyendo así a un ejercicio más riguroso y ético de la profesión.

Agradecimientos

Expresamos nuestro más sincero agradecimiento a los integrantes de la Comisión de Test y al personal administrativo del COP, así como a las personas que revisaron los test, a las casas editoriales y a Ismael Loinaz, autor de correspondencia del test no comercializado CTS2. Su ayuda, colaboración y cooperación fueron esenciales para llevar a cabo la undécima evaluación de test editados en España.

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. <https://www.papelesdelpsicologo.es/pdf/3033.pdf>
- American Educational Research Association, American Psychological Association, y National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Psychological Association.
- Arribas, D., Corral, S., y Pereña, J. (2022). *CTC-R: Cuestionario TEA Clínico - Revisado*. Hogrefe TEA Ediciones.
- Chapman, H., y Gillespie, S. M. (2019). The Revised Conflict Tactics Scales (CTS2): A review of the properties, reliability, and validity of the CTS2 as a measure of partner abuse in community and clinical samples. *Aggression and Violent Behavior*, 44, 27-35. <https://doi.org/10.1016/j.avb.2018.10.006>
- Corral, S., y Calvete, E. (2006). Evaluación de la violencia en las relaciones de pareja mediante las Escalas de tácticas para conflictos: estructura factorial y diferencias de género en jóvenes. *Psicología Conductual*, 14, 215-233.
- Departamento de I+D de Pearson Clinical & Talent Assessment, Hernández, A., Paradell, È., y Vallar, F. (2022). *BESS: Sistema de cribado conductual y emocional del BASC-3*. Pearson.
- Echeburúa, E., Amor, P. J., Loinaz, I., y Corral, P. (2010). Escala de predicción del riesgo de violencia grave contra la pareja-revisada (EPV-R). *Psicothema*, 22(4), 1054-1060. <https://www.psicothema.com/pdf/3840.pdf>
- Elosua, P., y Geisinger, K. F. (2016). Cuarta evaluación de tests editados en España: Forma y fondo. *Papeles del Psicólogo*, 37(2), 82-88. <https://www.papelesdelpsicologo.es/pdf/2693.pdf>
- Equipo R&D - Giunti Psychometrics España, Martínez, A., Miralles, J., y de Ancos, I. (2021). *SRP-4: Escala de psicopatía - 4ª Edición*. Giunti Psychometrics.
- Equipo R&D - Giunti Psychometrics España, Martínez, A., Miralles, J., y de Ancos, I. (2022). *MASC2: Escala de Ansiedad Multidimensional para Niños - 2ª Edición*. Giunti Psychometrics.
- Espejo-Navarro, A. L., y Valdivia-Ramírez, D. M. (2023). *Propiedades psicométricas de la Conflict Tactics Scale en estudiantes universitarios de Lima*. Universidad Peruana de Ciencias Aplicadas. <https://repositorioacademico.upc.edu.pe/handle/10757/658452>
- Evers, A., Muñoz, J., Hagemester, C., Høstmælingen, A., Lindley, P., Sjöberg, A., y Bartram, B. (2013). Assessing the quality of test: Revision of the EFPA review model. *Psicothema*, 25(3), 283-291. <https://doi.org/10.7334/psicothema2013.97>
- Fernández-Ballesteros, R., De Bruyn, E. E. J., Godoy, A., Hornke, L. F., Ter Laak, J., Vizcarro, C., ... y Zaccagnini, J. L. (2001). Guidelines for the assessment process (GAP): A proposal for discussion. *European Journal of Psychological Assessment*, 17(3), 187-200. <https://doi.org/10.1027//1015-5759.17.3.187>
- Fonseca, E., y Muñoz, J. (2017). Quinta evaluación de tests editados en España: mirando hacia atrás, construyendo el futuro. *Papeles del Psicólogo*, 38(3), 161-168. <https://doi.org/10.23923/pap.psicol2017.2844>
- Gallego Rodríguez, C., y Fernández-González, L. (2019). ¿Se relaciona el consumo de pornografía con la violencia hacia la pareja?: El papel moderador de las actitudes hacia la mujer y la violencia. *Psicología Conductual*, 27(3), 431-454.
- Galve Manzano, J. L., y Martínez Arias, R. (2019). *BECOLE-R: Baterías de Evaluación Cognitiva de las Dificultades en la Lectura y Escritura. Renovado y Revisado*. Editorial CEPE.
- Gómez Sánchez, 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>
- Graña, J. L., Andreu, J. M., Peña, M., y Rodríguez-Biezma, M. J. (2013). Validez factorial y fiabilidad de la "Escala de tácticas para el conflicto revisada" (Revised Conflict Tactics Scales, CTS2) en población adulta española. *Psicología Conductual*, 21(3), 525-543.
- Hernández, A., Elosua, P., Fernández-Hermida, J. R., y Muñoz, J. (2022). Comisión de Test: Veinticinco años velando por la calidad de los test. *Papeles del Psicólogo*, 43(1), 55-62. <https://doi.org/10.23923/pap.psicol.2978>
- 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. <https://doi.org/10.7334/psicothema2019.306>
- Hernández, A., Ponsoda, V., Muñoz, 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. <https://www.papelesdelpsicologo.es/pdf/2775.pdf>
- Hernández, A., Tomás, I., Ferreres, A., y Lloret, S. (2015). Tercera evaluación de tests editados en España. *Papeles del Psicólogo*, 36(1), 1-8. <https://www.papelesdelpsicologo.es/pdf/2484.pdf>
- Hidalgo, M. D., y Hernández, A. (2019). Sexta evaluación de test editados en España: resultados e impacto del modelo en docentes y editoriales. *Papeles del Psicólogo*, 40(1), 21-30. <https://doi.org/10.23923/pap.psicol2019.2886>
- International Test Commission (2013). *ITC guidelines on test use*. https://www.intestcom.org/files/guideline_test_use.pdf
- International Test Commission (2014). *ITC statement on the use of tests and other assessment instruments for research purposes*. https://www.intestcom.org/files/statement_using_tests_for_research.pdf
- International Test Commission. (2017). *The ITC guidelines for translating and adapting tests (Second edition)*. https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf
- International Test Commission y Association of Test Publishers (2022). *Guidelines for technology-based assessment*. <https://www.intestcom.org/upload/media-library/guidelines-for-technology-based-assessment-v20221108-16684036687NAG8.pdf>
- Loinaz, I. (2009). *Aproximación teórica y empírica al estudio de las tipologías de agresores de pareja. Análisis descriptivo y variables e instrumentos de evaluación en el centro penitenciario Brians-2*. Ministerio del Interior, Secretaría General Técnica.
- Loinaz, I., Echeburúa, E., Ortiz-Tallo, M., y Amor, P. J. (2012). Propiedades psicométricas de la Conflict Tactics Scales (CTS-2) en una muestra española de agresores de pareja. *Psicothema*, 24(1), 142-148. <https://www.psicothema.com/pi?pii=3991>
- 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., Onís, M. C., Dueñas, H. R., Albert, C. C., Aguado, T. C., y Luque, L. R. (2002). Versión española del cuestionario de Yesavage abreviado (GDS) para el despistaje de depresión en mayores

- de 65 años: adaptación y validación. *Revista de Medicina Familiar y Comunitaria*, 12, 620-630.
- Martínez de la Iglesia, J., Onís, M. C., Dueñas, H. R., Albert, C. C., Aguado, T. C., Colomer, A., ... y Blanco, M. C. (2005). Abreviar lo breve. Aproximación a versiones ultracortas del cuestionario de Yesavage para el cribado de la depresión. *Atención Primaria*, 35(1), 14-21. <https://doi.org/10.1157/13071040>
- Muñiz, J., Elosua, P., y Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema*, 25(2), 151-157. <https://doi.org/10.7334/psicothema2013.24>
- 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. <https://www.papelesdelpsicologo.es/pdf/1947.pdf>
- 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>
- Muñoz-Sánchez, S. Y. (2018). *Violencia de pareja y resolución de conflictos en relaciones LGTBI en Bogotá*. Universidad Nacional de Colombia. <https://repositorio.unal.edu.co/handle/unal/69077>
- Redondo, N., y Graña, J. I. (2015). Consumo de alcohol, sustancias ilegales y violencia hacia la pareja en una muestra de maltratadores en tratamiento psicológico. *Adicciones*, 27(1), 27-36. <https://doi.org/10.20882/adicciones.191>
- 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>
- Portellano, J. A., Mateos, R., Martínez-Arias, R., y Sánchez-Sánchez, F. (2021). *CUMANIN-2: Cuestionario de Madurez Neuropsicológica Infantil-2*. Hogrefe TEA Ediciones.
- Prieto, G., y Muñiz, J. (2000). Un modelo para evaluar la calidad de los test utilizados en España. *Papeles del Psicólogo*, 77, 65-71. <https://www.papelesdelpsicologo.es/resumen?pii=1102>
- Schittekatte, M., y Evans, N. (2023, September 27). Updating the EFPA BoA Test Review Model: a necessary titanic work with many angles and supported by even more shoulders. Comunicación oral presentada en el simposio "Updates on the work of the EFPA Board of Assessment" del 18th European Congress of Psychology, Brighton (United Kingdom). <https://doi.org/10.23668/psycharchives.13272>
- Straus, M. A., Hamby, S. L., Boney-McCoy, S. U. E., y Sugarman, D. B. (1996). The revised conflict tactics scales (CTS2) development and preliminary psychometric data. *Journal of Family Issues*, 17(3), 283-316. <https://doi.org/10.1177/019251396017003001>
- Toro, J., y Cervera, M. (1984). *T.A.L.E.: Test de análisis de lectoescritura*. Machado Grupo de Distribución, S.L.
- 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. <https://doi.org/10.23923/pap.psicol2020.2937>