GRUPO EDITORIAL DE PSICOFUNDACIÓN

Papeles del Psicólogo (2024) 45(2) 56-64

Papeles del Psicólogo
Psychologist Papers

https://www.papelesdelpsicologo.es/ • ISSN 0214–7823 • eISSN 1886–1415

PAPELES DEL PSICÓLOGO
PSYCHOLOGIST PAPERS

Revista del Consejo General de la Psicología de España

Artículo

Décima Evaluación de Test Editados en España: Incorporando Información Sobre Test no Comerciales

Francisco José Abad

Universidad Autónoma de Madrid, España

INFORMACIÓN

Recibido: Enero 10, 2024
Aceptado: Marzo 4, 2024

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

RESUMEN

Los test son fundamentales para la Psicología, y su uso debe estar siempre respaldado por evidencias sólidas. Desde 2010, la Comisión de Test del Consejo General de la Psicología en España (COP) realiza evaluaciones anuales, utilizando el Cuestionario de Evaluación de Test Revisado (CET-R) y contando con la colaboración de expertos externos. Hasta la fecha, se han revisado 96 test. Esta décima edición incluye seis test de tres casas editoriales reconocidas como TEA Hogrefe, Pearson Educación y Giunti Psychometrics. Además, se revisan las propiedades psicométricas de una prueba no comercial, la Escala de Depresión Geriátrica (GDS), mencionada entre los 25 test más usados por psicólogos españoles. Evaluar test no comerciales, desarrollados en contextos académicos, es crucial, ya que enriquece el conjunto de herramientas disponibles para los profesionales. Este trabajo aborda los desafíos de evaluar pruebas de este tipo y ofrece sugerencias para mejorar tanto el desarrollo y validación de los test como la revisión de sus propiedades psicométricas mediante el CET-R.

Tenth Review of Tests Published in Spain: Incorporating Information on non-Commercial Tests

Keywords
Tests
Assessing test quality
Psychometrics
Psychometric properties
CET-R

ABSTRACT

Tests are fundamental to psychology, and their use should always be supported by solid evidence. Since 2010, the National Test Commission of the General Council of the Spanish Psychological Association has been carrying out annual evaluations, using the Test Evaluation Questionnaire-Revised (CET-R) with the collaboration of external experts. To date, 96 tests have been evaluated. This tenth edition includes six tests from three well-known publishing houses: TEA Hogrefe, Pearson Educación, and Giunti Psychometrics. In addition, the psychometric properties were also reviewed of a non-commercial test, the Geriatric Depression Scale (GDS), mentioned among the 25 most used tests by Spanish psychologists. Evaluating non-commercial tests, developed in academic contexts, is crucial, as it enriches the set of tools available to practitioners. This paper addresses the challenges of evaluating tests of this type and offers suggestions for improving both the development and validation of the tests as well as the review of their psychometric properties using the CET-R.

Cómo citar: Abad, F. J. (2024). Décima evaluación de test editados en España: incorporando información sobre test no comerciales. *Papeles del Psicólogo/Psychologist Papers, 45*(2), 56-64. https://doi.org/10.23923/pap.psicol.3033
Autor de correspondencia: fjose.abad@uam.es

La última encuesta a psicólogos españoles sobre test (Muñiz et al., 2020) muestra su uso habitual en la práctica profesional, considerándose de gran ayuda para la toma de decisiones en ámbitos muy diversos (educativo, organizacional, clínico y de la salud, etc.). Los test son reconocidos por su amplia gama de funciones, tales como el diagnóstico, la selección, la orientación y la adaptación de intervenciones. Estas herramientas se distinguen por su naturaleza estandarizada y propiedades psicométricas conocidas, lo que contribuye a mejorar la precisión en la predicción y el diagnóstico. No obstante, dada la vasta cantidad de test disponibles y las consecuencias que pueden tener las puntuaciones en determinados contextos, es crucial que los profesionales puedan elegir de manera informada. En este sentido, se hace esencial que cada uso específico de un test esté respaldado por la evidencia.

En esta línea, numerosas asociaciones y organismos han desarrollado iniciativas para garantizar la calidad y buen uso de los test. Entre estas, destacan los "Estándares para Test Psicológicos y Educativos", establecidos por las principales asociaciones americanas de Psicología y Educación (AERA et al., 2014). Además, la Comisión Internacional de Test (International Test Commission, ITC) ha publicado diversas directrices para guiar sobre aspectos clave de la construcción y aplicación de test, tales como la traducción y adaptación (ITC, 2018; Muñiz et al., 2013), la seguridad (ITC, 2014) o el uso de las tecnologías en la evaluación (ITC, 2022).

Por otro lado, se han diseñado acciones específicas para la revisión y evaluación de la calidad de los test psicológicos y educativos. En EE.UU., el sistema BUROS (Carlson y Geisinger, 2012) es un notable ejemplo, mientras que en Europa se dispone del modelo de revisión de test de la Federación Europea de Asociaciones de Psicología (EFPA; Evers et al., 2013), que está en proceso de actualización (Schittekatte et al., 2023). En España, la Comisión de Test del Consejo General de la Psicología en España (COP) lleva publicando evaluaciones desde 2011 (Muñiz et al., 2011). Las cuatro primeras ediciones (1ª, Muñiz et al., 2011; 2ª, Ponsoda y Hontangas, 2013; 3ª, Hernández et al., 2015; 4ª, Elosua y Geisenger, 2016) se hicieron siguiendo una versión adaptada del modelo europeo, el Cuestionario de Evaluación de Test (CET; Prieto y Muñiz, 2000). Para las ediciones posteriores (5ª, Fonseca-Pedrero y Muñiz, 2017; 6ª, Hidalgo y Hernández, 2019; 7ª, Gómez, 2019; 8ª, Viladrich et al., 2021; 9ª, Lozano, 2023) se ha utilizado una nueva versión del cuestionario CET (CET-R; Hernández et al., 2016), que recoge las actualizaciones del modelo europeo (Evers et al., 2013). Hasta la fecha se han revisado 96 test, como se detalla en la Tabla 1. El número de test distintos es algo menor, ya que algunas pruebas han requerido múltiples evaluaciones (p. ej., para la BADYG se realiza un informe por nivel) o han sido objeto de revisión en más de una ocasión (p. ej., PAIB).

La Tabla 1 muestra que las revisiones se han centrado principalmente en test comerciales, lo que concuerda con los resultados de la encuesta de Muñiz et al. (2020), donde la mayoría de los test más utilizados por psicólogos españoles, excepto la Escala de Depresión Geriátrica (GDS), son comerciales. Hasta la fecha, solo se ha analizado un test no comercial, la escala revisada de predicción del riesgo de violencia grave contra la pareja (EPV-R; Echeburúa et al., 2010; revisada en Ponsoda y Hontangas, 2013). Sin embargo, muchos test de fuentes académicas se usan tanto en la academia como fuera de ella (ver, por ejemplo, el Banco de Instrumentos y Metodologías en Salud Mental, CIBERSAM, s.f., que lista pruebas como el SDQ, aplicado en la Encuesta Nacional de Salud, Ortuño et al., 2016). Por ello, la Comisión de Test se propuso avanzar en la evaluación de pruebas no comerciales usadas profesionalmente. En la presente edición se ha evaluado la GDS (Yesavage y Sheikh, 1986) y está previsto que en la siguiente se evalúen las Escalas de Tácticas de Conflicto (CTS; Straus, 1979). Nótese que no es la primera vez que se utiliza el CET-R como marco de referencia para evaluar las propiedades psicométricas de este tipo de test. Más allá de la evaluación de la EPV-R, por el COP, también se ha utilizado el CET-R para la revisión de test de Regulación Emocional (Pérez-Sánchez et al., 2020, 2022) o para valorar puntualmente alguna propiedad psicométrica de un cuestionario (p.ej., Espada et al., 2022; Reyes-Pérez et al., 2023).

Revisar pruebas no comerciales conlleva afrontar desafíos significativos, como la falta de manuales completos y la dispersión de la información relevante. En el caso de adaptaciones, deben establecerse criterios claros para determinar qué versiones de un test evaluar. Para seleccionar de entre la diversidad de adaptaciones, deben considerarse factores como la fidelidad de la adaptación al

Tabla 1

Listado de Test Evaluados por el COP en Cada Edición

Año[1]	Coordinador/a	Test	Instrumentos
2024	F. J. Abad	7	**BASC-3**, CIT, CPF, CTE, DABS, **GDS**, PROLEXIA
2023	L. Lozano	6	BAYLEY-III, BECOLE-R, CAG, DAS, MacArthur, **Raven 2**
2021	C. Viladrich	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
2019	L. Gómez	8	BRIEF-2, BYI-2, DP-3, Factor-g-R, IAES-A, *PAIB-1*, *PAIB-2*, *PAIB-3*
2019	M.D. Hidalgo	10	*BADyG/E2-r*, *BADyG/S-r*, BAT-7, BDI-FastScreen, BPR, CESPRO, MATRICES, MBMD, Perfil Sensorial-2, Q-PAD
2017	E. Fonseca	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**
2016	P. Elosua	11	ABAS-II, *BADyG/M-r*, BETA, BSI-18, CECAD, EHPAP, *PAIB* (*1*), PECC, SCIP-S, WMS-IV, WPPSI-IV
2015	A. Hernández	11	BCSE, BECOLE, Boehm-3, Boehm-3 Preescolar, CESQT, ECLE, ESQUIZO-Q, IECI, SOC, TRauma, ***WAIS-IV***
2013	V. Ponsoda	12	BAI, BAS-II, BDI-II, CEAM, CompeTEA, ESCOLA, ESPERI, EPV-R, MPR, PAI, RIAS, WNV
2011	J. Muñiz	8	EFAI, EVALUA, IGF, **MMPI-2-RF**, **NEO PI-R**, **PROLEC-R**, ***WISC-IV*** , **16PF-5**

Notas. En negrita aparecen los que, en alguna versión, han sido citados entre los 25 test más utilizados por los psicólogos españoles (Muñiz et al., 2020). En cursiva, los test que han sido evaluados en una segunda ocasión o constituyen una versión más reciente de la misma batería (la versión más reciente aparece subrayada); [1]Año de publicación del trabajo describiendo el proceso de revisión de la correspondiente edición.

instrumento original, el uso extendido en la práctica clínica y en investigaciones previas, así como la calidad de la traducción y adaptación cultural. Es igualmente importante considerar la información disponible sobre las propiedades psicométricas de la adaptación en muestras locales. En el caso de pruebas desarrolladas originalmente en español, será igualmente relevante determinar cómo seleccionar los estudios más pertinentes.

El rol del coordinador en esta evaluación incluye algunas funciones adicionales, siendo responsable de seleccionar las fuentes documentales adecuadas, lo que implica una búsqueda meticulosa en bases de datos académicas y la revisión de literatura relevante. Esto incluye estudios de validación, como análisis factoriales y correlaciones con otras escalas estandarizadas, y de baremación, aunque estos últimos son menos frecuentes. Además, se requiere limitar el número de artículos para la evaluación, incluyendo todos los estudios relevantes de la versión adaptada, pero estableciendo un límite razonable con relación al número de artículos sobre el cuestionario original. En casos específicos, también puede ser útil consultar a expertos en el campo o a los autores de la prueba para acceder a documentación adicional sobre las propiedades psicométricas de la prueba. En los casos de aplicación del CET-R a la evaluación de pruebas no comerciales existe variedad de aproximaciones a la selección de la documentación; desde casos en que la documentación ha sido limitada y propuesta por los autores de la escala (caso del EPV-R, revisado en Ponsoda y Hontangas, 2013), hasta casos en que la evaluación se ha basado en una búsqueda intensiva en bases de datos, considerando conjuntamente estudios de validación sobre la escala en distintos países (p. ej., Pérez-Sánchez et al., 2020).

Método

Participantes

Para esta edición se contactó con 15 expertos, de los cuales uno declinó participar por encontrarse ya jubilado. En la Tabla 2 se muestran las 14 personas que finalmente han participado en esta edición (i.e., dos revisores por test) y que fueron seleccionadas procurando mantener los criterios de paridad de sexo y diversidad geográfica. En la mayor parte de los casos cada test fue revisado por un experto con perfil metodológico y otro experto en la variable medida por el test. Por lo tanto, hay profesores de las áreas de Metodología de las Ciencias del Comportamiento, de Personalidad, Evaluación y Tratamientos Psicológicos y de Psicología Evolutiva y de la Educación, así como algunos especialistas clínicos. También se tuvo en cuenta que no hubiera conflicto de intereses ni relación directa con los autores.

Tabla 2
Revisores Participantes en la Décima Evaluación de Test

Nombre	Filiación
Juan Ramón Barrada González	Universidad de Zaragoza
Paula Elosua Oliden	Universidad del País Vasco
Sergio Escorial Martín	Universidad Complutense de Madrid
David Gallardo-Pujol	Universitat de Barcelona
Eduardo García-Garzón	Universidad Camilo José Cela
Ana Hernández-Baeza	Universidad de Valencia
Alicia Eva López Martínez	Universidad de Málaga
Estela López Nicolás	Centro de Intervención Psicoeducativa Huarte de San Juan (Navarra)
Fabia Morales Vives	Universitat Rovira i Virgili
Amparo Oliver Germes	Universidad de Valencia
Mireia Orgilés Amorós	Universidad Miguel Hernández
Patricia Recio Saboya	UNED
Francisco J. Román González	Universidad Autónoma de Madrid
Miguel Angel Sorrel Luján	Universidad Autónoma de Madrid

Instrumento

Para llevar a cabo la evaluación de los test se utilizó el Cuestionario para la Evaluación de los Test Revisado (CET-R; Hernández et al., 2016). El CET-R consta de unas breves instrucciones en las que se ofrecen al revisor algunas observaciones generales e importantes, tales como no considerar información ajena a la documentación entregada o, en el caso de pruebas adaptadas, ponderar de forma distinta los estudios de la versión original y la forma adaptada. Por otro lado, se ofrece un glosario de términos psicométricos, para facilitar que todos los revisores les asignen el mismo significado. El cuestionario consta de tres secciones principales: Descripción general del test, Valoración de las características del test y Valoración global del test.

La primera sección, Descripción general del test, contiene 28 ítems que proporcionan información editorial y de autoría del test (p. ej., fechas de publicación del test original y de la adaptación, precio de manual y cuadernillos), así como información sobre los constructos medidos, el diseño del test (p. ej., número de escalas/ítems, formato de respuesta, soporte, tiempo de aplicación) y otros aspectos de uso (p. ej., áreas de aplicación, poblaciones a las que va dirigido, cualificación requerida para su uso).

En la segunda sección, Valoración de las características del test, se describen las propiedades psicométricas de las puntuaciones e incluye cuatro apartados:

a) Valoración general del test: 10 ítems que evalúan la calidad de la fundamentación teórica, el proceso de desarrollo y análisis de los ítems, instrucciones, así como los materiales y documentación aportada;
b) Validez: 19 ítems que valoran las evidencias de validez relacionada con el contenido, estructura interna y relación de las puntuaciones del test con otras variables, entre otros aspectos;
c) Fiabilidad: 14 ítems para evaluar aspectos como la equivalencia entre formas paralelas, la estabilidad test-retest, la consistencia interna, la fiabilidad inter-jueces y, cuando se aplica la teoría de la respuesta al ítem (TRI), la función de información; y
d) Baremos e interpretación de puntuaciones: 9 ítems centrados en la calidad del proceso de baremación para interpretación normativa y de los puntos de corte para la interpretación referida a criterio.

Mediante las preguntas se valoran tanto los índices psicométricos obtenidos (p. ej., la cuantía de los coeficientes de fiabilidad) como la calidad de los estudios que los respaldan. Se consideran tanto aspectos cuantitativos (p. ej., el número de expertos en los estudios de validez de contenido, el tamaño muestral en los de consistencia interna) como cualitativos (p. ej., la calidad de los criterios empleados en la validez referida a criterio o la adecuación del baremo a la población objetivo). Por último, tres subsecciones (validez, fiabilidad, baremos e interpretación de puntuaciones) finalizan con una sección abierta de comentarios en la cual el revisor debe resumir y justificar las puntuaciones asignadas.

Tabla 3
Listado de Test Evaluados en la Décima Edición

Acrónimo	Nombre	Autor/es originales (año de publicación)	Autor/es adaptación (año de publicación)	Editorial
BASC-3	Sistema de evaluación de la conducta de niños y adolescentes	Reynolds y Kamphaus (2015)	Departamento de I+D de Pearson Clinical & Talent Assessment: Ana Hernández, Èrica Paradell y Frédérique Vallar (2020)	Pearson Educación
CIT	Cuestionario de Impacto del Trauma	--	Crespo, González-Ordi, Gómez-Gutiérrez y Santamaría (2020)	TEA Hogrefe
CPF	Cuestionario de Personalidad Forense	--	Medina y Sintas (2021)	Giunti Psychometrics
CTE	Cuestionario de Talento Emprendedor	--	Valderrama (2021)	Giunti Psychometrics
DABS	Escala de Diagnóstico de Conducta Adaptativa	Tassé, Schalock, Balboni, Bersani, Borthwick-Duffy, Spreat, Thissen, Widaman y Zhang (2017)	Verdugo, Arias y Navas (2021)	TEA Hogrefe
GDS	Escala de Depresión Geriátrica	Yesavage y Sheikh (1986)	Martínez de la Iglesia, Onís, Dueñas, Albert, Aguado y Luque (2002)	---
PROLEXIA	Diagnóstico y Detección Temprana de la Dislexia	--	Cuetos, Arribas, Suárez-Coalla y Martínez-García (2020)	TEA Hogrefe

La tercera sección (Valoración general) recoge un resumen cuantitativo (promedios) de los resultados de la sección anterior y una parte abierta en la que deben reflejarse las fortalezas y debilidades de la prueba, sugerencias de uso para profesionales y recomendaciones para la mejora de la prueba. En la versión final del informe, publicada en la web del COP, esta valoración se presenta al inicio. En la puntuación final asignada a cada test se agregan los ítems cuantitativos de cada sección (33): Materiales y documentación (2 ítems), Fundamentación teórica (1 ítem), Adaptación (1 ítem), Análisis de los ítems (1 ítem), Validez: contenido (2 ítems), Validez: relación con otras variables (9 ítems), Validez: estructura interna (1 ítem), Validez: análisis del funcionamiento diferencial de los ítems (1 ítem), Fiabilidad: equivalencia (3 ítems), Fiabilidad: consistencia interna (2 ítems), Fiabilidad: estabilidad (2 ítems), Fiabilidad: TRI (2 ítems), Fiabilidad: inter-jueces (1 ítem) y Baremos e interpretación de las puntuaciones (5 ítems).

Los ítems adoptan generalmente, el siguiente sistema de etiquetas: 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' incluye una descripción detallada para guiar a los evaluadores sobre lo que representa esta puntuación en cada apartado. Además, en aquellos ítems donde es posible una cuantificación más objetiva, se utilizan etiquetas específicas para facilitar una valoración más precisa. Por ejemplo, en la evaluación de la consistencia interna, se utiliza una escala en la que 3 = *Adecuada* ($0.70 \leq r < 0.80$). En situaciones donde el manual no provee la información necesaria para responder a un ítem, algunas preguntas permiten distinguir entre casos en los que la característica o sección no es aplicable al instrumento (donde no se otorga puntuación) y aquellos en los que, siendo aplicable, falta la información requerida (en cuyo caso se asigna una puntuación de cero). El CET-R está disponible para su consulta y descarga en la página web del COP (https://www.cop.es/test/).

Procedimiento

Como en anteriores revisiones, las editoriales (Giunti Psychometrics, Pearson Educación y TEA-Hogrefe) propusieron a la Comisión de Test del COP las pruebas que querían someter a evaluación (i.e., seis test). La selección de test a evaluar desde las editoriales se hizo en dos tandas (los primeros cuatro test se enviaron a revisores en julio de 2021 y el resto en octubre). Adicionalmente, desde la Comisión de Test, se decidió añadir una prueba no comercializada, la GDS (Yesavage et al., 1982; Yesavage y Sheikh, 1986). Los test evaluados se muestran en la Tabla 3.

El proceso de revisión para las pruebas comerciales siguió un protocolo similar al de ediciones anteriores. En cada caso, el coordinador se puso en contacto con los revisores y, una vez aceptaban, el editor proporcionaba una copia completa de cada prueba al coordinador y a cada revisor. Adicionalmente, el coordinador enviaba a los revisores el CET-R, dando un plazo amplio de cuatro meses para su cumplimentación. Las respuestas de revisión por pares al CET-R se recibieron hasta abril de 2022. En marzo y abril, el coordinador preparó un informe provisional para cada prueba, integrando las valoraciones de ambos expertos. No siempre hubo concordancia entre las calificaciones, pero en caso de discrepancias, el coordinador determinaba la calificación final, teniendo en cuenta el razonamiento del experto y la información del manual de la prueba. En mayo el informe preliminar se envió a las casas editoriales, que tuvieron el plazo de un mes para presentar alegaciones. Finalmente, el coordinador preparó la versión definitiva del informe, atendiendo a la información proporcionada por los expertos y las casas editoriales. Estos informes definitivos fueron revisados por un miembro de la Comisión de test del COP, cuyas sugerencias de estilo fueron incorporadas antes del envío final para su publicación (en septiembre de 2022).

Protocolo Específico con Respecto a la Escala no Comercial, la GDS

Con relación al GDS, se generó un listado de las versiones adaptadas, haciendo una búsqueda bibliográfica en la Web of Science[1] y partiendo del trabajo de revisión de Cabañero-Martinez et al. (2007). Con esos criterios se localizaron un total de 103 artículos y más de 20 versiones que variaban en número de ítems (siendo 15 y 30 las longitudes más frecuentes). En segundo lugar, se establecieron los criterios para elegir la versión a revisar. Se consideraron: (a) número de citas recibidas en Google Scholar, Web of Science y/o Scopus; (b) inclusión en trabajos de revisión de las

1 Términos de búsqueda: ("Geriatric depression scale" OR GDS* OR Yesavage) (Topic) and (spanish OR Spain) (All Fields) and (GDS* OR Yesavage OR depression OR depresión) (Title)

adaptaciones de escalas de depresión o en bancos de instrumentos de salud mental (p. ej., CIBERSAM); (c) reconocimiento por parte del autor original de la adaptación de la escala (p. ej., listando la versión en su página web); (d) tamaño y representatividad de las muestras de validación.

A partir de la información recogida se valoró que la versión abreviada de Martínez de la Iglesia et al. (2002, 2005), de 15 ítems, parecía ser la opción más popular, atendiendo, por ejemplo, al número de citas de los trabajos, a la inclusión en la revisión de instrumentos de screening para la depresión, en español, de Reuland et al. (2009) o en la página web del autor de la versión original. Además, dicha versión es una de las dos mejor puntuadas en la revisión de Cabañero-Martínez et al. (2007), que consideran que es la única en la que se informa de la realización de un adecuado proceso de adaptación transcultural. En cuanto al cuarto criterio considerado, las propiedades psicométricas de esta versión se estudiaron originalmente en una muestra de tamaño mayor que el de otras adaptaciones (ver Tabla 4) y un estudio reciente aporta información para la obtención de datos normativos (Delgado-Losada et al., 2021). Otro punto positivo a tener en cuenta es la brevedad de esta versión, en comparación con las formas de 30 ítems, lo que ayuda a disminuir los problemas de fatiga y falta de atención que comúnmente ocurren en el grupo etario al que la escala va dirigida. Por último, un análisis de los enunciados de los ítems mostró que su contenido era fiel al de la versión original, frente a otras versiones en las que se hacen modificaciones importantes (p.ej., la de Ortega-Orcos et al., 2007).

Tras seleccionar la versión específica de la GDS de Martínez de la Iglesia et al. (2002, 2005), se realizó una revisión exhaustiva de los artículos que citaban estos trabajos, utilizando las bases de datos Scopus y Web of Science (WoS). Se localizaron 224 trabajos. Finalmente, se decidió seleccionar un conjunto de 14 artículos entre los que cabe mencionar los tres artículos fundamentales sobre el desarrollo de la versión original de la escala (Brink et al., 1982; Yesavage et al., 1982, 1986), siete artículos que proporcionaban información relativa a las propiedades psicométricas de la versión adaptada (los más importantes: Martínez de la Iglesia et al., 2002, 2005; Lucas-Carrasco, 2012; Delgado-Losada et al., 2021), y cuatro trabajos de revisión o síntesis tanto de la versión adaptada (Cabañero-Martínez et al., 2007) como de la original (p. ej., Balsamo et al., 2018). Adicionalmente, se proporcionó a los revisores un compendio de 55 artículos citados en estas revisiones para permitirles profundizar en aspectos más específicos si lo consideraban necesario; entre ellos cabría citar el funcionamiento diferencial de los ítems, estudios sobre la estructura interna, o investigaciones acerca de la generalización de la fiabilidad de la escala.

Tabla 4

Distintas Versiones de las Escalas GDS-15 y GDS-30 en Español (Selección de Artículos con más de 20 Citas en Google Scholar[5] y Muestras de más de 50 Evaluados)

Autores	Año	Ítems	Google scholar	N
Abizanda et al.[1]	1998	30	32	142
De Dios et al.[1]	2001	15	40	155
Fernández-San Martin et al.[1,2,3]	2002	30	129	192
García-Serrano y Tobías[1]	2001	30	112	173
Izal y Montorio[1]	1993	30	63	60
Martí et al[1]	2000	15	64	131
Martinez de la Iglesia et al.[1,2,4]	2002	15	187	249
Ortega-Orcos et al.[2,3]	2007	15	38	301
Salamero y Marcos[1]	1992	30	95	234

Nota. [1]Citado en Cabañero et al. (2017); [2]Citado en Reuland et al., 2009; [3]Citado en Mitchell et al., 2010; [4]Adaptación española citada en la web del autor original: https://web.stanford.edu/~yesavage/GDS.html; [5] actualizado el 12/2023

Resultados

Los informes detallados correspondientes a los test evaluados en esta décima edición se pueden consultar y descargar en la página web del COP, dentro del apartado correspondiente al año 2021 (https://www.cop.es/test/#evaluados). En nuestro caso, la mediana de los coeficientes de correlación entre las puntuaciones otorgadas por el Revisor 1 y el Revisor 2 en las preguntas en que ambos dieron puntuaciones válidas fue de 0,59, similar a la publicada por Ponsoda y Hontangas (2013) que fue de 0,61. Este nivel medio-bajo de acuerdo no es atípico de este tipo de evaluaciones (Hogan et al., 2021). En el apartado de discusión se recogen algunas posibles razones para los desacuerdos.

En la Tabla 5 se muestra un resumen de las puntuaciones obtenidas en cada apartado para cada uno de los 7 test. Como se puede observar, el patrón de resultados es muy similar al obtenido en ediciones previas.

El primer bloque de puntuaciones se refiere a la consideración de Aspectos generales y de desarrollo de la prueba. En los apartados de Materiales y documentación y de Fundamentación teórica, las pruebas comerciales obtienen puntuaciones que, en general, pueden calificarse de buenas o excelentes (promedios de 4,5 y 4,2, respectivamente). Para el GDS, el primer apartado no pudo ser evaluado, dada la ausencia de manual o cuadernillos impresos, pero obtiene una puntuación excelente (4,5) en el apartado de fundamentación teórica. El proceso de Adaptación se considera bueno o excelente en todas las adaptaciones.

El segundo bloque recoge las valoraciones de las evidencias de validez de las pruebas. Con relación a las evidencias de validez de contenido, los test comerciales reciben, en general, calificaciones buenas o excelentes, indicando que se cuida este aspecto (p.ej., con una buena fundamentación teórica, a través de la consulta a expertos, revisiones de los ítems y/o la realización de estudios piloto). No obstante, no en todos los casos se proporciona información cuantitativa y detallada de este proceso, lo cual podría ser deseable. Para el GDS se obtiene una puntuación más baja, ya que los revisores consideran que no se ha obtenido evidencia de este tipo.

Con relación a la evidencia de relación con otras variables, las puntuaciones son entre adecuadas y excelentes, y en promedio son buenas (promedio = 3,7). Es relevante destacar que varios de los test incluyen criterios para valorar la sensibilidad y la especificidad de los puntos de corte propuestos, lo que permite ir más allá de la interpretación normativa.

Con relación a las evidencias relativas a la estructura interna, la puntuación promedio está por debajo de lo encontrado en ediciones anteriores (3,1 vs. 3,8). Esto se debe a que tres escalas han recibido una puntuación más baja (2 = *Adecuado con carencias*), explicada por distintas razones según cada caso, como la falta de evidencia para algunas escalas, la incompleta provisión de información o la evidencia desfavorable para algunas de las escalas. En el caso del GDS es necesario ampliar el número de estudios que empleen muestras locales. Con relación al análisis del funcionamiento diferencial de los ítems (DIF) es de valorar que cada vez sea más

Tabla 5
Puntuaciones Obtenidas por los Test Analizados en la Décima Evaluación

	BASC3	CIT	CPF	CTE	DABS	GDS	PROLEXIA	Promedio	Histórico*
Desarrollo: Materiales y documentación	4,8	5	3,5	3,8	5	--	5	4,5	4,3
Desarrollo: Fundamentación teórica	4	5	2,5	3,5	5	4,5	5	4,2	4,1
Desarrollo: Adaptación	3,5	--	--	--	5	4,5	--	4,3	4,3
Desarrollo: Análisis de los ítems	--	4	2,5	3,5	4,5	3	5	3,8	3,8
Validez: contenido	4	5	-	3,5	5	1,5	4	3,8	3,8
Validez: relación con otras variables	3	4,8	3,5	2,8	3,7	4,3	4	3,7	3,6
Validez: estructura interna	2	4,5	3	2	4,5	2	3,5	3,1	3,8
Validez: análisis del DIF	--	4	--	--	3	4	--	3,7	--
Fiabilidad: equivalencia	--	--	--	--	--	--	--	--	--
Fiabilidad: consistencia interna	4,5	4,5	3	2,5	4,5	4,3	4,5	4,0	4,2
Fiabilidad: estabilidad	3,5	4	3,5	--	3	2,5	3,5	3,3	3,5
Fiabilidad: TRI	--	--	3,5	--	4	4,5	--	4,0	--
Fiabilidad: inter-jueces	--	--	--	--	5	3	--	4,0	--
Baremos e interpretación de las puntuaciones	4	4,3	3,8	3,2	4,3	4	4,5	4,0	4,1

Nota. Las puntuaciones de la tabla siguen una escala de 1 a 5: 1 = Inadecuado; 2 = Adecuado con carencias; a partir de 2,5, Adecuado; a partir de 3,5, Bueno; a partir de 4,5 = Excelente. El símbolo -- indica que no se aporta información o no procede; *Puntuación media en las ediciones anteriores.

frecuente la comprobación de la invarianza de las puntuaciones a través de los grupos (p.ej., edad y sexo), lo que es fundamental para garantizar la equidad de la evaluación.

El tercer bloque recoge evidencia sobre la precisión de las pruebas. Como en ediciones previas, se encuentra que la fiabilidad se valora principalmente mediante indicadores de consistencia interna. Las valoraciones son en su mayoría buenas o excelentes (el valor más bajo, 2,5, es adecuado), lo que implica el uso de muestras de tamaño suficiente y valores de consistencia interna aceptables. No obstante, cabe destacar como limitación que algunas de las pruebas no incluyen los indicadores de consistencia interna de todas las escalas, lo que debería cuidarse en futuras ediciones de sus manuales. Con respecto a la estabilidad, los valores son algo inferiores, pero pueden considerarse adecuados para todas las pruebas (salvo para el GDS, lo que se debe a la escasez de estudios con muestra local, que se penaliza). Para los casos en los que se aplica la TRI o se calcula la fiabilidad inter-jueces se encuentran puntuaciones promedio buenas.

El último bloque recoge las valoraciones de la calidad de los Baremos e interpretación de las puntuaciones. En promedio, se obtienen también puntuaciones buenas, pero en el caso de dos escalas, se obtienen calificaciones por debajo de lo deseado. Una limitación importante es que no se ha recogido evidencia para su aplicación en algunas de las muestras objetivo para las que se propone su uso. Entre los apartados analizados destaca la actualización de las normas y el uso de la tipificación continua (continuous norming), que optimiza la eficiencia en el proceso de construcción de los baremos, especialmente en aquellos casos en los que se trabaja con población infantil y/o adolescente y los constructos evaluados siguen una tendencia a través de la edad (Evers et al., 2010; Evers et al., 2013).

Conclusiones

Valoración Global y Posibles Mejoras

Los resultados indican la alta calidad de los test editados en España, con buenas puntuaciones y estudios exhaustivos de las propiedades psicométricas, incluyendo el uso de técnicas avanzadas como la TRI o la tipificación continua. Es valorable que varias pruebas incluyan estudios sobre los puntos de corte, que enriquecen la interpretación de las puntuaciones. No obstante, se recomienda incrementar los estudios sobre el funcionamiento diferencial de los ítems y precisar las hipótesis al describir las evidencias de validez convergente y discriminante. También es fundamental evitar la extrapolación de resultados de estudios de validación de una muestra a otra con características distintas. Por último, se observa que cada vez es más frecuente omitir los baremos del manual, lo que dificulta la evaluación de estos. Nuestra recomendación es que las editoriales proporcionen la información como material adicional para el proceso de revisión del test.

Evaluación de Pruebas no Comerciales

La evaluación de la prueba no comercial implicó retos específicos: (a) Algunos criterios importantes del CET-R resultan inaplicables a pruebas carentes de manual ni materiales impresos, lo que resalta la necesidad de adaptar el enfoque de evaluación, pero también nos lleva a recomendar a los investigadores que generen este tipo de materiales; (b) Equilibrar la carga de trabajo para los revisores con el acceso a la información relevante fue complejo. La aproximación de hacer una revisión sistemática/meta-análisis de todas las publicaciones hacía inviable la revisión y la inclusión única de los trabajos sobre la versión adaptada se podía quedar corta, por lo que se optó por una postura intermedia, proporcionando como documentación los principales trabajos de validación de la escala en muestra local, a la vez que incluyendo trabajos más generales de revisión. Nos encontramos que los revisores tendían a complementar la documentación, por lo que la labor de coordinación e integración fue mayor en este caso; (c) La heterogeneidad de la calidad de los estudios de validación puede ser mayor que en el caso de test comerciales, por lo que supone un factor adicional de complejidad; (d) Para algunos criterios, una mayoría de estudios se refería a la versión original, pero el CET-R no especifica cuánto peso se debe dar a estos estudios; (e) La ausencia de baremos en publicaciones científicas puede complicar la evaluación; en la elección de la versión a analizar, la existencia de un artículo reciente con normas fue decisiva.

Con Relación al CET-R y Posibles Mejoras de Este

En nuestra experiencia con el CET-R, identificamos algunos problemas, varios de ellos mencionados en ediciones previas. Primero, no todos los revisores sumaban las puntuaciones de los ítems para obtener las globales, a pesar de las instrucciones claras. Este problema se resolvería proporcionando una plantilla de cálculo automático a los evaluadores o digitalizando el CET-R para su uso en línea.

Segundo, hubo inconsistencia en el manejo de puntuaciones de cero cuando faltaba información en el manual. Algunos revisores las incluían en los promedios, mientras que otros no. Esta discrepancia puede deberse a la ambigüedad en las instrucciones del CET-R, que sugiere promediar solo los apartados con información disponible. Nuestra revisión de evaluaciones anteriores mostró que generalmente no se asignaban puntuaciones de cero ni se consideraban en los promedios, por lo que seguimos esa práctica.

En tercer lugar, el acuerdo entre los revisores varió según los criterios, siendo menor para algunos criterios específicos (p. ej., evidencias de estructura interna, calidad de los baremos). Las discrepancias pueden deberse a múltiples razones, algunas de las cuales ya han sido señaladas en revisiones previas. Una primera razón es que los expertos en contenido y en metodología son sensibles a distintos aspectos, siendo los últimos más exigentes en la aplicación de los procedimientos. En otros casos, las discrepancias pueden deberse a las dificultades de valoración en casos complejos. Por ejemplo, la no inclusión de baremos en el manual era muy penalizada por algunos revisores, pero no por otros. La valoración de los tamaños muestrales también puede ser compleja, cuando el test tiene distintas versiones que se aplican en distintas muestras o cuando se utiliza la tipificación continua. Por último, los revisores varían en el grado de penalización cuando el manual no proporciona información relevante para una o más escalas, cuando la información se refiere a las versiones originales de la escala, o cuando se propone el uso del test en varias poblaciones, pero solo se proporcionan normas apropiadas para una de ellas. Una breve guía de ejemplos comentados podría ser de utilidad para homogeneizar y facilitar las valoraciones.

CET-R v1.1

Existe ya una versión revisada del CET-R, en la que se han implementado varias mejoras importantes. En la nueva versión se distingue más claramente entre la información esencial para evaluar la calidad de una prueba que no se presenta y aquella que, aunque ausente, no es fundamental para su propósito. Además, para las pruebas adaptadas, se solicita a los revisores que especifiquen la procedencia de las muestras en los distintos apartados (Análisis de ítems, Validez, etc.), permitiendo así evaluar el grado de validación de la prueba con muestras locales. También se subraya la necesidad de utilizar muestras locales para que el baremo sea adecuado. Por último, se incorporan orientaciones para la valoración del Área Bajo la Curva (AUC) en el uso de curvas ROC, un aspecto cada vez más relevante en estudios de sensibilidad y especificidad de un test, especialmente en la predicción de criterios específicos, como categorías diagnósticas.

Conclusiones Finales

Para concluir, es valioso reflexionar sobre el impacto del proceso de revisión de test. Se ha observado un efecto positivo general, particularmente en la presentación más detallada de las evidencias que respaldan la calidad técnica de los test en los manuales recientes. El modelo CET-R no solo guía a autores y editores en el desarrollo y adaptación de pruebas, sino que también contribuye a la difusión de pruebas menos conocidas pero ampliamente utilizadas y respaldadas por evidencias de calidad técnica y psicométrica. Además, es una herramienta formativa importante para los futuros psicólogos, concienciándolos sobre los estándares que deben exigir en la aplicación de los test.

Respecto a cómo mejorar el conocimiento de estos procesos entre los psicólogos españoles, sería conveniente crear una base de datos de test evaluados, organizada por constructos o áreas de evaluación, facilitando así comparativas. El COP ya ha dado un paso en esta dirección con su Buscador de test (https://www.jornadas.cop.es/evaluacionTest/), que permite búsquedas por palabras clave y mejora el acceso a estas evaluaciones. Además, es crucial evaluar cómo los psicólogos utilizan estos informes y su utilidad práctica, así como entender los criterios que emplean al elegir un test.

Por último, es esencial perseverar en la revisión de test no comerciales que se usan profesionalmente, a pesar de las dificultades inherentes a este proceso. Dicha evaluación es clave ya que si la evaluación es favorable enriquece el conjunto de herramientas disponibles para los profesionales, mientras que si es desfavorable mitiga los riesgos de uso de test inadecuados, basados en normas obsoletas o validados en muestras no adecuadas.

Agradecimientos

Agradezco la colaboración de los miembros de la Comisión de Test, en particular a Ana Hernández y Paula Elosua, por su valiosa ayuda y cooperación en todo el proceso. También quiero expresar mi gratitud a las casas editoriales por proporcionar los ejemplares a evaluar, así como por brindar retroalimentación detallada y constructiva en sus alegaciones a los informes provisionales. Por último, reconocer y agradecer a los revisores por su generosa y excelente colaboración, sin la cual estas evaluaciones no serían posibles.

Conflicto de Intereses

No existe conflicto de intereses.

Referencias

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

Abizanda, P., Luengo, C., López, J., Sánchez, P., Romero, L., y Fernández, C. (1998). Predictores de mortalidad, deterioro funcional e ingreso hospitalario en una muestra de ancianos residentes en la comunidad. *Revista Española de Geriatría y Gerontología*, *33*, 219-225.

Balsamo, M., Cataldi, F., Carlucci, L., Padulo, C., y Fairfield, B. (2018). Assessment of late-life depression via self-report measures: A review. *Clinical Interventions in Aging*, *13*, 2021-2044. https://doi.org/10.2147/CIA.S178943

Brink, T. L., Yesavage, J. A., Lum, O., Heersema, P. H., Adey, M., y Rose, T. L. (1982). Screening Tests for Geriatric Depression. *Clinical Gerontologist*, *1*(1), 37-43. https://doi.org/10.1300/J018v01n01_06

Cabañero-Martínez, M. J., Richart-Martínez, M., Muñoz-Mendoza, C. L., y Reig-Ferrer, A. (2007). Revisión estructurada de las escalas de depresión en personas mayores. *International Journal of Clinical and Health Psychology*, *7*(3), 823-846.

Carlson, J. F., y Geisinger, K. F. (2012). Test reviewing at the Buros Center for Testing. *International Journal of Testing*, *12*, 122-135. https://doi.org/10.1080/15305058.2012.661003

Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM) (s.f.). *Banco de Instrumentos y Metodologías en Salud Mental*. Ministerio de Ciencia e Innovación. https://bi.cibersam.es/busqueda-de-instrumentos

Crespo, M., González-Ordi, H., Gómez-Gutiérrez, M., y Santamaría, P. (2020). *CIT: Cuestionario de Impacto del Trauma*. Hogrefe TEA Ediciones.

Cuetos, F., Arribas, D., Suárez-Coalla, P., y Martínez-García, C. (2020). *PROLEXIA. Diagnóstico y Detección Temprana de la Dislexia*. Hogrefe TEA Ediciones.

Delgado-Losada, M. L., López-Higes, R., Rubio-Valdehita, S., Facal, D., Lojo-Seoane, C., Montenegro-Peña, M., Frades-Payo, B., y Fernández-Blázquez, M. Á. (2021). Spanish consortium for ageing normative data (SCAND): Screening tests (MMSE, GDS-15 and MFE). *Psicothema*, *33*(1), 70-76. https://doi.org/10.7334/psicothema2020.304

Dios, R. de, Hernández, A. M., Rexach, L. I., y Cruz, A. J. (2001). Validación de una versión de cinco ítems de la Escala de Depresión Geriátrica de Yesavage en una población española. *Revista Española de Geriatría y Gerontología*, *36*, 276-280. https://doi.org/10.1016/S0211-139X(01)74736-1

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.

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

Espada Sánchez, J. P., Gonzálvez Maestre, M. T., Fernández Martínez, I., Orgilés Amorós, M., y Morales Sabuco, A. (2022). Spanish validation of the short mood and feelings questionnaire (SMFQ) in children aged 8-12. *Psicothema*, *34*(4), 610-620. https://doi.org/10.7334/psicothema2022.54

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. https://psycnet.apa.org/doi/10.1080/15305058.2010.518325

Evers, A., Muñiz, J., Hagemeister, 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. https://doi.org/10.7334/psicothema2013.97

Fernández-San Martín, M. I., Andrade, C., Molina, J., Muñoz, P. E., Carretero, B., Rodríguez, M., y Silva, A. (2002). Validation of the Spanish version of the geriatric depression scale (GDS) in primary care. International. *Journal of Geriatric Psychiatry*, *17*(3), 279-287. https://psycnet.apa.org/doi/10.1002/gps.588

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

García-Serrano, M. J., y Tobías, J. (2001). Prevalencia de depresión en mayores de 65 años. Perfil del anciano de riesgo. *Atención Primaria*, *27*, 484-488. https://doi.org/10.1016/S0212-6567(01)78839-7

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

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G., y Elosua, P. (2016). Revisión del modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo/Psychologist Papers*, *37*(1), 192-197. https://www.papelesdelpsicologo.es/pdf/2775.pdf

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

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

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

International Test Commission (2014). International Guidelines on the Security of Tests, Examinations, and Other Assessments. https://www.intestcom.org/files/guideline_test_security.pdf

International Test Commission and Association of Test Publishers (2022). *Guidelines for technology based assessment*. https://www.intestcom.org/page/28 and https://www.testpublishers.org/white-papers

International Test Commission. (2018). ITC Guidelines for Translating and Adapting Tests. *International Journal of Testing*, *18*(2), 101-134. https://doi.org/10.1080/15305058.2017.1398166

Izal, M., y Montorio, I. (1993). Adaptation of the Geriatric Depression Scale: A preliminary study. *Clinical Gerontologist*, *13*, 83-91. https://doi.org/10.1300/J018v13n02_07

Lozano, L. M. (2023). Novena evaluación de los test editados en España. *Papeles del Psicólogo/Psychologist Papers*, *44*(1), 1-7. https://doi.org/10.23923/pap.psicol.3004

Lucas-Carrasco, R. (2012). Reliability and validity of the Spanish version of the World Health Organization-Five Well-being Index in elderly. *Psychiatry and Clinical Neurosciences*, *66*(6), 508-513. https://doi.org/10.1111/j.1440-1819.2012.02387.x

Martí, D., Miralles, R., Llorach, I., García-Palleiro, P., Esperanza, A., Guillén, J., y Cervera, A. (2000). Trastornos depresivos en una unidad de convalecencia: experiencia y validación de una versión española de 15 preguntas de la escala de depresión geriátrica de Yesavage. *Revista Española de Geriatría y Gerontología*, *35*, 1-7.

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., Arias, C., y Blanco, M. C. (2005). Abbreviating the brief. Approach to ultra-short versions of the Yesavage questionnaire for the diagnosis of depression. *Atención Primaria*, *35*(1), 14-21. https://doi.org/10.1157/13071040

Medina, P. M., y Sintas, F. (2021). *Cuestionario de Personalidad Forense.* Giunti EOS Psychometrics.

Mitchell, A., Bird, V., Rizzo, M., y Meader, N. (2010). Diagnostic validity and added value of the geriatric depression scale for depression in primary care: A meta-analysis of GDS(30) and GDS(15). *Journal of Affective Dissorders*, *125*(1-3), 10-17. https://doi.org/10.1016/j.jad.2009.08.019

Muñiz, J., Elosua, P., y Hambleton, y R. K. (2013). Directrices para la traducción y adaptación de los test: 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, Á., y Peña-Suárez, E. (2011). Evaluación de test editados en España. *Papeles del Psicólogo/Psychologist Papers*, *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/Psychologist Papers*, *41*(1), 1-15. https://doi.org/10.23923/pap.psicol2020.2921

Ortega Orcos, R., Fort, M. S., Khajoui, A. K., Aparicio, S. V., y Valle, R. D. D. del (2007). Validación de la versión española de 5 y 15 ítems de la Escala de Depresión Geriátrica en personas mayores en Atención Primaria. *Revista Clínica Española*, *207*(11), 559-562. https://doi.org/10.1016/S0014-2565(07)73477-X

Ortuño-Sierra, J., Fonseca-Pedrero, E., Inchausti, F., y Sastre i Riba, S. (2016). Evaluación de dificultades emocionales y comportamentales en población infanto-juvenil: El cuestionario de capacidades y dificultades (SDQ). *Papeles del psicólogo/Psychologist Papers*, *37*(1), 14-26. https://www.papelesdelpsicologo.es/pdf/2658.pdf

Pérez-Sánchez, J., Delgado, A. R., y Prieto, G. (2020). Psychometric properties of the scores of the most commonly used tests in the evaluation of emotion regulation. *Papeles del Psicólogo/Psychologist Papers*, *41*(2), 116-124. https://doi.org/10.23923/pap.psicol2020.2931

Pérez-Sánchez, J., Delgado, A. R., y Prieto, G. (2022). Evaluación del Emotion Regulation Checklist para Niños y Adolescentes. *Psicologia: Teoria e Pesquisa*, *38*. https://doi.org/10.1590/0102.3772e38213.es

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

Prieto, G., y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo/Psychologist Papers*, *77*, 65-77. https://www.papelesdelpsicologo.es/resumen?pii=1102

Reuland, D. S., Cherrington, A., Watkins, G. S., Bradford, D. W., Blanco, R. A., y Gaynes, B. N. (2009). Diagnostic accuracy of Spanish language depression-screening instruments. *Annals of Family Medicine*, *7*(5), 455-462. https://doi.org/10.1370/afm.981

Reyes-Pérez, Á., López-Martínez, A. E., Esteve, R., y Ramírez-Maestre, C. (2023). Spanish Validation of the COMM Scale to Assess the Misuse of Prescription Opioids in Patients with Chronic Noncancer Pain. *International Journal of Mental Health and Addiction*, *21*(5), 3458-3472. https://doi.org/10.1007/s11469-022-00803-3

Reynolds, C. R., y Kamphaus, R. W. (2015). *BASC3: Behavior Assessment System for Children*. Pearson.

Salamero, M., y Marcos, T. (1992). Factor study of the Geriatric Depression Scale. *Acta Psychiatrica Scandinavica*, *86*, 283-286.

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. [Conference object] Symposium at European Congress of Psychology 2023 Brighton. https://doi.org/10.23668/psycharchives.13272

Straus, M. A. (1979). Measuring intrafamily conflict and violence: The Conflict Tactics Scales. *Journal of Marriage and Family*, *41*, 75-88. https://psycnet.apa.org/doi/10.2307/351733

Tassé, M. J., Schalock, R. L., Balboni, G., Bersani, H., Borthwick-Duffy, S. A., Spreat, S., Thissen, D., Widaman, K. F., y Zhang, D. (2017). *Diagnostic Adaptive Behavior Scale (DABS) User's Manual*. American Association on Intelectual and Developmental Disabilities.

Valderrama, B. (2021). *CTE. Cuestionario de Talento Emprendedor.* Giunti EOS Psychometrics.

Verdugo, M. A., Arias, B., y Navas, P. (2021). *Escala de Diagnóstico de Conducta Adaptativa (DABS).* Hogrefe TEA Ediciones.

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/Psychologist Papers*, *42*(1), 1-9. https://doi.org/10.23923/pap.psicol2020.2937

Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., y Leirer, V. O. (1982). Development and validation of a geriatric depression screening scale: A preliminary report. *Journal of Psychiatric Research*, *17*(1), 37-49. https://doi.org/10.1016/0022-3956(82)90033-4

Yesavage, J. A., y Sheikh, J. I. (1986). 9/Geriatric Depression Scale (GDS): Recent Evidence and Development of a Shorter Version. *Clinical Gerontologist*, *5*(1-2), 165-173. https://doi.org/10.1300/J018v05n01_09