GRUPO EDITORIAL DE PSICOFUNDACIÓN

Papeles del Psicólogo (2025) 46(3) 158-166

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

Article

Eleventh Review of Tests Published in Spain

Georgina Guillera & Maite Barrios

University of Barcelona, Spain

ARTICLE INFO

Received: January 22, 2025
Accepted: April 30, 2025

Keywords
Tests
Test evaluation
Psychometrics
Psychometric properties
CET-R

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 the Spanish Psychological Association 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 ones 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 their technical updates and strengthening transparency and confidence in the instruments used by psychologists.

Undécima Evaluación de Test Editados en España

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.

Cite this article as: Guillera, G,. & Barrios, M. (2025). Eleventh Review of Tests Published in Spain. *Papeles del Psicólogo/Psychologist Papers, 46*(3), 158-166.
https://doi.org/10.70478/pap.psicol.2025.46.20
Correspondence: Georgina Guilera gguilera@ub.edu

Introduction

Psychological tests are fundamental tools in the professional practice of psychology, frequently used to support decisions in various fields such as clinical, educational, organizational, and health psychology. These tests allow us to diagnose, select, guide, and adapt interventions, and their value lies in their standardized nature and the strength of their psychometric properties, which provide accuracy and confidence in the results (Muñiz et al., 2020). However, due to the large number of tests available and the important consequences that can result from their applications, it is essential that psychologists have access to verified and detailed information on the quality of these instruments in order to choose the most appropriate ones in each assessment context.

In response to this need, international bodies, such as the American Psychological Association (APA), together with the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME), have developed rigorous standards to ensure the quality and appropriate use of tests (AERA, APA, & NCME, 2014). Others, such as the International Test Commission (ITC), have developed guidelines on test translation and adaptation (Hernández et al., 2020; ITC, 2017; Muñiz et al., 2013), and on test use in general (ITC, 2013) and in specific domains (e.g., research: ITC, 2014; digital environments: ITC and Association of Test Publishers, 2022). Regarding test quality assessment, in Europe, the European Federation of Psychologists' Associations (EFPA) has developed and implemented a comprehensive evaluation model to assess the validity and reliability of the inferences drawn from tests used in the region (Evers et al., 2013), which is currently being updated (Schittekatte & Evans, 2023). In the Spanish context, the Test Commission of the General Council of the Spanish Psychological Association (COP) has taken the lead in the systematic review of tests published in Spain since 2010 (Hernández et al., 2022). This initiative arose in response to the demand of psychologists for independent evaluations that provide technical and practical information on the quality of the instruments they use in their daily practice (Muñiz et al., 2011; Hernández et al., 2016). To carry out these reviews, the Test Evaluation Questionnaire (CET in Spanish) has been used, which was initially developed by Prieto and Muñiz (2000) and has been modified on several occasions to align with European models and to incorporate the most current technological advances and psychometric analysis practices (Hernández et al., 2016).

To date, multiple editions of the test review have been carried out in Spain, each of which has resulted in the publication of detailed reports accessible to professionals through the COP website (https://www.cop.es/test/). The importance of these reviews is evident in the results of studies such as those of Muñiz et al. (2020), which show a growing use of tests by Spanish psychologists, who positively value their usefulness for decision making in diverse contexts. However, they also stress the need for more technical information to enable them to make evidence-based decisions on the suitability of tests for specific uses. In this regard, only 22.5% of the licensed psychologists surveyed in the study by Muñiz et al. stated that they were aware of the annual evaluation carried out by the Test Commission, although those who were aware considered these evaluations to be useful and necessary, and stated that they consult the reports to decide which tests to use in their professional practice (Muñiz et al., 2020). Beyond the benefits that this evaluation process offers to psychologists applying tests, it is also desirable that it contributes to improving the quality of the tests themselves—encouraging ongoing development in their construction, updating, and application.

This manuscript presents, firstly, the results from the eleventh edition of the evaluation of tests published in Spain, conducted between 2022 and 2024. Secondly, it presents the findings from a comparative analysis of the information reported and the psychometric quality of the tests before and after the implementation of the test evaluation system.

Eleventh Test Review

This section outlines the tests that were evaluated, the evaluation procedure followed, the individuals who participated as reviewers, and the main results of the eleventh review of tests published in Spain.

Tests Submitted for Evaluation

Following the procedure adopted in previous editions, the publishers represented on the Test Commission (i.e., Editorial CEPE, Giunti Psychometrics, Hogrefe TEA Ediciones, and Pearson Clinical Assessment Spain) proposed to the commission the six tests they wished to submit for evaluation. Additionally, the Test Commission agreed to include a seventh test that has not been commercialized in Spain, the revised version of the Conflict Tactics Scales (CTS-2; Straus et al., 1996), adapted to Spanish by Loinaz (2009) and validated in the work of Loinaz and colleagues (Loinaz et al., 2012).

In accordance with this approach, seven tests have been evaluated in this eleventh edition, covering different areas of applied psychology, such as clinical, educational, forensic, and work and organizational psychology, as well as other areas such as neuropsychology or research. Table 1 lists the seven instruments submitted for review, six of which are commercially available through publishing companies.

Selection of Reviewers

Building on previous editions of the test evaluation process, and with the aim of assigning to each test one reviewer with expertise in technical and psychometric aspects and another with experience in the variables assessed by the tests, potential reviewers were sought by combining several search strategies. For the identification of experts in psychometrics, (a) the lists of reviewers of the ten previous editions of the test review published in *Papeles del Psicólogo* were consulted (Abad, 2024; Elosua & Geisinger, 2016; Fonseca & Muñiz, 2017; Gómez-Sánchez, 2019; Hernández et al., 2015; Hidalgo & Hernández, 2019; Lozano, 2023; Muñiz et al., 2011; Ponsoda & Hontangas, 2013; Viladrich et al., 2021); (b) contact information was extracted for members of the European Association of Methodology (EAM) with Spanish affiliation (download date from the association's website: June 30, 2022); and, finally, (c) the list was supplemented by other researchers with expertise in psychometrics and methodology, identified within the coordinators'

Table 1
List of the Measuring Instruments of the Eleventh Edition of the Test Review

Acronym	Test	Publisher	Author(s)	Year of publication/update
BESS	*Sistema de cribado conductual y emocional del BASC-3* [BASC-3 behavioral and emotional screening system].	Pearson Education	Pearson Clinical & Talent Assessment R&D Department, A. Hernández, È. Paradell, & F. Vallar	2022
CTC-R	*Cuestionario TEA Clínico - Revisado* [Clinical TEA Questionnaire - Revised]	Hogrefe TEA Editions	D. Arribas, S. Corral, & J. Pereña	2022
CTS-2	*Versión revisada de la Conflict Tactics Scales* [Revised version of the Conflict Tactics Scales]	--	I. Loinaz, E. Echeburúa, M. Ortiz-Tallo, & P. J. Amor	2012
CUMANIN-2	*Cuestionario de madurez neuropsicológica infantil-2* [Child Neuropsychological Maturity Questionnaire-2]	Hogrefe TEA Editions	J. A. Portellano, R. Mateos, R. Martínez-Arias, F. Sánchez-Sánchez	2021
MASC2	*Escala de Ansiedad Multidimensional para Niños/as* [Multidimensional Anxiety Scale for Children]	Giunti Psychometrics Spain	R&D Team - Giunti Psychometrics Spain, A. Martínez, J. Miralles, & I. de Ancos	2022
SRP 4	*Escala de Psicopatía* [Psychopathy Scale]	Giunti Psychometrics Spain	R&D Team - Giunti Psychometrics Spain, A. Martínez, J. Miralles, & I. de Ancos	2021
T.A.L.E.	*Test de análisis de lectoescritura* [Literacy analysis test]	Machado Grupo de Distribución, S.L. [Machado Distribution Group, a Spanish limited liability company]	J. Toro & M. Cervera	1984

network of contacts. In the case of the experts in the construct, for their identification, (a) the lists of reviewers of previous editions were also consulted; and (b) searches were made in electronic databases using, on the one hand, the name of the test and, on the other, the construct measured by each of the tests. An initial group of 14 reviewers was contacted (i.e., 7 psychometricians and 7 construct experts), of whom 28.6% (i.e., 4 psychometricians) accepted the invitation in the first round; vacancies were subsequently filled with new reviewers. A second round of invitations was issued to cover the remaining positions, which enabled full coverage of psychometric experts for all tests. However, multiple attempts were needed to secure the participation of construct experts in the review process. Specifically, whereas 11 researchers were contacted to cover the psychometrician profile, 17 were contacted to fill the profile of construct experts. Table 2 presents the list of reviewers who participated in the eleventh test review.

Table 2
List of Reviewers of the Eleventh test Review

Name and surname	Affiliation
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)

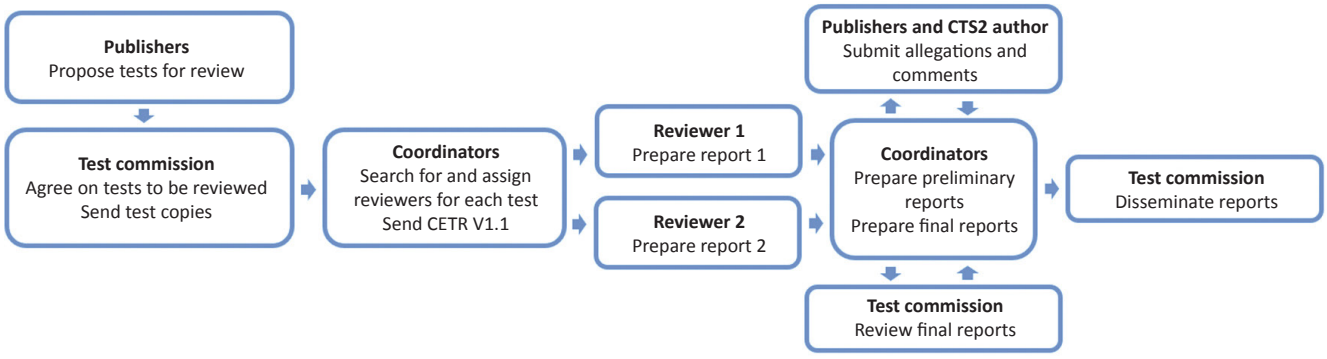
Evaluation Instrument: The CET-R V1.1

For the evaluation of the tests, the Revised Test Evaluation Questionnaire (CET-R; Hernández et al., 2016), specifically designed to describe and evaluate the technical and psychometric quality of tests, was used in its most current version (i.e., version V1.1, available on the COP website: https://www.cop.es/test/). Abad's work (2024) describes in detail the structure of the instrument in sections (i.e., General description of the test, Assessment of test characteristics, and Global assessment of the test), the aspects of the test evaluated in each section—either by means of open or closed questions—as well as the scoring system and the corresponding labels of the quantitative items. In this latest version of the instrument, which was already used in Abad (2024), significant improvements have been introduced, such as (1) a clearer distinction between information that is essential for assessing the quality of a test and that which, although absent, is not essential given the purpose of the test; (2) more specific criteria for tests adapted from other languages and/or cultures, in which reviewers are asked to indicate the origin of the samples (i.e., local, international, or mixed), (3) guidelines for assessing the Area Under the Curve (AUC) in the use of ROC curves, a key aspect in the prediction of diagnostic criteria.

Evaluation Process

With regard to the six tests that are commercially available, once the reviewers agreed to participate, the administrative staff of the COP coordinated the distribution of the test materials to

Figure 1
Test Review Procedure Followed in the Eleventh Edition



each of the two reviewers, as well as to those responsible for coordinating the evaluation process. At the same time, the coordinator of the Test Committee sent the contact details of the representatives of the publishers and the CET-R V1.1 to the coordinators of this edition of the test review. In the case of the Spanish version of the CTS-2, since it is not commercially available, a multi-step literature search was conducted to locate relevant materials. First, studies citing the Spanish validation study (Loinaz et al., 2012) were identified in Scopus (n = 43) and Web of Science (n = 45) [search completed in July 2022]. Second, publications with the term "Conflict Tactics Scales-2" in the title, abstract, or keywords were searched in Scopus (n = 63) and Web of Science (n = 18) [also in July 2022]. After reviewing the titles and abstracts—and, when necessary, the full texts—publications that the coordinators judged to contain relevant information for assessing the test's quality were retained (e.g., empirical studies and reviews on the psychometric properties of the scale). Third, from examining the list of references of the selected studies, seven additional relevant papers were identified (i.e., Corral & Calvete, 2006; Espejo-Navarro & Valdivia-Ramírez, 2023; Gallego Rodríguez & Fernández-González, 2019; Graña et al., 2013; Loinaz, 2009; Muñoz-Sánchez, 2018; Redondo & Graña, 2015). This literature search process ended with the creation of a list of 50 papers, mostly articles that employed the Spanish version of the STS-2, although it also included articles on the original English version (e.g., Straus et al., 1996), as well as reviews of the psychometric properties of the scale (e.g., Chapman & Gillespie, 2019). This documentation was provided to the two reviewers assigned to evaluate the STS-2.

After receipt of the documentation, the reviewers proceeded to evaluate the test independently through the CET-R V1.1. Reviewer reports were submitted through April 2023. For each test, the coordinators synthesized the evaluations of the two assigned reviewers and prepared preliminary reports, which were sent to the publishers and the corresponding author of the validation study of the Spanish version of the CTS-2 to provide them with the opportunity to present arguments. Finally, the coordinators prepared the final reports, which were reviewed by a member of the Test Commission. Their suggestions, mainly stylistic and clarifying, were incorporated for the publication of the final version of the reports on the COP website in May 2024. Figure 1 shows an outline of the test review procedure followed in the eleventh edition.

Results

Table 3 shows the scores resulting from the review process for the various tests together with the average score for all the characteristics assessed. In general, all mean scores reflect good to excellent ratings for the most recently published or updated commercial tests. In contrast, the CTS-2—a non-commercial test—and the T.A.L.E.—which has not been updated in the last four decades—show inadequate scores or, at best, adequate but with shortcomings.

Comparative Analysis of Data Reporting and Test Quality Before and After the Start of the Evaluation Process

Objective

As noted, the review process for tests published in Spain, led by the COP Test Commission, was established to address one of the main demands of professional psychologists: to provide technical information that facilitates informed decision making. After completing eleven review processes, we proposed the possibility of comparatively analyzing the information report and the quality of the tests published before and after the start of the review system, taking into account the results of having applied the CET, in its different versions, throughout the eleven editions. The underlying assumption is that the introduction of the review system should have had an impact at two levels. Firstly, it was expected that more information would be reported about the psychometric properties of the tests. Secondly, the tests themselves would receive improved ratings, given that test developers are expected—albeit gradually—to take into account the standards and recommendations put forward by the Test Commission. Specifically, this study explores to what extent the implementation of the review system has led to an increase in (1) the amount of information provided regarding the psychometric characteristics of the tests (the quantity of information), and (2) the scores on each of the test characteristics evaluated (the quality of the tests).

Procedure

Scores were extracted from the 96 reports available on the Test Commission website from the ten editions completed to date, to which were added those corresponding to the seven tests of the

Table 3
Scores of the Tests Evaluated in the Eleventh Edition

Characteristics	BESS	CTC-R	CTS-2	CUMANIN-2	MASC2	SRP 4	T.A.L.E.
Development							
Materials and documentation	4.5	5	-	5	4.5	3.8	4
Theoretical foundation	4	5	3	4	4	5	1
Adaptation	4	-	3	-	3	3	-
Item analysis	-	4	4	5	4	-	2
Validity							
Content	4	4	2.5	4.5	3	3.5	1.5
Relationship with other variables	3.9	4.4	2.6	4.2	4	3.8	-
Internal structure	2	4.5	2	5	4	3	-
DIF Analysis	-	5	-	5	5	-	-
Reliability							
Equivalence	-	-	-	-	-	-	-
Internal consistency	3.3	5	3	5	4	3.8	-
Stability	3.5	3.5	2	4.3	3.5	2.3	-
IRT	-	-	-	3	4	-	-
Inter-rater	-	-	-	-	-	-	-
Scales and interpretation of scores	4	5	3	5	3.5	3	1.7
Total	3.7	4.5	2.8	4.5	3.9	3.5	2.0

Note: DIF: differential item functioning; IRT: item response theory.The scores correspond to a five-point rating scale, where: 1 = Inadequate, 2 = Adequate with some deficiencies, 3 = Adequate, 4 = Good, and 5 = Excellent.The symbol (-) indicates that no information is provided or it is not applicable due to the characteristics of the test.

present edition, resulting in a total of 103 reports. The three tests that are not commercially available were excluded from the analysis: the *Escala de Predicción del riesgo de Violencia grave contra la pareja-Revisada* [Severe Intimate Partner Violence Risk Prediction Scale-Revised] (EPV-R; Echeburúa et al., 2010), the Geriatric Depression Scale - short version (GDS; Martínez de la Iglesia et al., 2002, 2005), and the Revised Conflict Tactics Scales (CTS-2; Loinaz, 2009; Loinaz et al., 2012).

The tests were classified as follows into three categories according to their year of publication or most recent update date: (1) the period prior to the start of the test review process by the Test Commission, which included tests published up to 2011, the year in which the results of the first edition were published (i.e., Muñiz et al., 2011); (2) the adaptation period, defined as a two-year time interval during which test developers had the opportunity to familiarize themselves with and adjust to the standards and criteria outlined in the evaluation model, which included tests published in 2012 and 2013; and (3) the period following the implementation of the test review process, which included tests published from 2014 onwards. This classification yielded data for tests published before (n = 39) and after (n = 50) the start of the test review process, the ratings of which were included in the comparative study (see Figure 2).

From each report, information was extracted from the general assessment of the respective tests (General assessment section of the CET V1.1), which includes 14 characteristics (see Table 3, Characteristics column). For each characteristic or psychometric property, two pieces of information were coded: first, whether the characteristic was reported or not (Reported vs. Not Reported), and second, the specific score assigned to that characteristic (numerical value from 0 to 5). Given that the CET model used across the successive editions has undergone some modifications, the 14 characteristics of the most current version of the CET-R (i.e., CET-R V1.1) were used as the baseline model. For most of the characteristics, there was a direct correspondence between the versions used in the different reports (e.g., evidence based on content). However, for some characteristics, a correspondence between reports had to be established (for example, "Factor analysis results" were categorized under "Validity: internal structure" and "Predictive validity" under "Validity: relationships with other variables"), which in some cases required the calculation of average scores.

In the comparison of reported information (i.e., Reported *vs.* Not reported) between the periods before and after the implementation of the review system, the χ^2 test was applied, accompanied by the ϕ coefficient as a measure of effect size. As regards the comparative study of test quality, the Mann-Whitney U test was applied to the scores received for each characteristic or psychometric property, and the effect size η^2 was calculated. The database and syntax used in this analysis can be found at the following link: https://osf.io/t43gf/.

Results

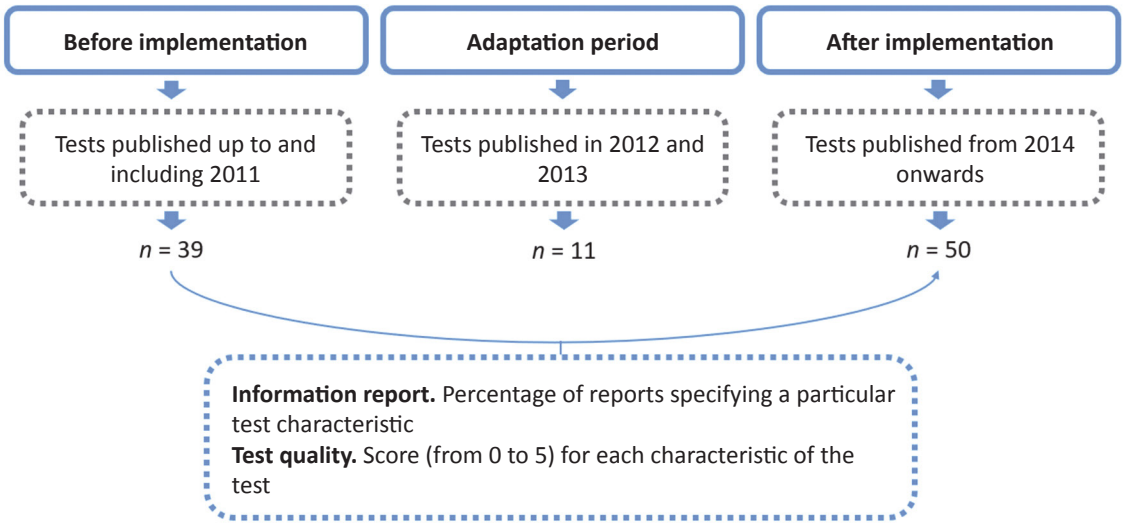
Information reporting. As shown in Table 4, it was observed that, regardless of the score obtained, the tests published after the introduction of the review process present a higher percentage of information regarding some psychometric properties (e.g., differential item functioning and temporal stability).

Quality of the tests. Table 4 presents the average scores obtained before and after the implementation of the review system, along with the value of the statistic and its corresponding effect size. Although for several of the characteristics analyzed the average score range was higher in the tests published after the implementation of the evaluation process, the Mann Whitney U test was not statistically significant in any case (see Table 4).

While this comparative analysis may encourage reflection on the test review process, it is not without limitations. One of the main challenges during the extraction of data from each of the test reports was the need to make adjustments due to changes in the different versions of the CET, which consequently affected how information was recorded in the reports. These changes required modifications

Figure 2

Outline of the Procedure Used in the Comparison Between the Periods Before and After the Implementation of the Test Review System



to extract the information and scores corresponding to the 14 characteristics of the General Assessment. For example, in the first reports for the first and second test evaluations, it was not possible to extract a score for 'Validity: Internal Structure'. Similarly, in the reports for the third and fourth round of reviews, the 'Construct Validity' section included analysis of internal structure, group comparison, convergent and discriminant evidence, among other aspects. However, in the most recent versions of the CET, this section was divided into 'Validity: Internal Structure' and 'Validity: Relationships to Other Variables', which made it necessary to assign this characteristic the score of 'Factor Analysis Results'. Additionally, some psychometric characteristics and analyses, such as 'Reliability: IRT' and 'Inter-rater reliability', were incorporated into the CET at later stages. Although it is possible that the description and assessment of these aspects were included in the open-ended sections of the reports, the present comparative analysis did not incorporate the information contained in such sections. Finally, the adaptation period was set at two years. However, this time may be insufficient for test developers to have become familiar with and adjust to the new requirements. These limitations highlight the need to consider possible biases or shortcomings when interpreting the results obtained in the present work.

Table 4

Comparison of Information Reporting and test Quality Between the Period Before (n = 39) and After (n = 50) the Implementation of the Test Review System

	Report (% Yes)					Quality (Average)				
Characteristics	**Before implementation**	**After implementation**	**χ2 (g.l.)**	***p* value**	**ϕ**	**Before implementation**	**After implementation**	***ZU***	***p* value**	**η2**
Development										
Materials and documentation	100	100	-	-	-	4.2	4.4	-1.364	.172	0.02
Theoretical foundation	100	100	-	-	-	4.1	4.3	-1.005	.315	0.01
Adaptation	100	100	-	-	-	4.3	4.3	-0.219	.827	< 0.01
Item analysis	79.5	84.0	0.303(1)	.582	0.06	3.9	3.8	-0.582	.561	< 0.01
Validity										
Content	94.9	94.0	0.031(1)	.895	-0.02	3.8	3.9	-0.139	.890	< 0.01
Relationship with other variables	97.4	100	1.297(1)	.255	0.12	3.7	3.7	-0.173	.862	< 0.01
Internal structure	84.6	84.0	0.006(1)	.937	-0.01	3.8	3.7	-0.019	.985	< 0.01
DIF Analysis	10.3	30.0	5.087(1)	.024	0.24	4.4	4.1	0.000	1.000	0.00
Reliability										
Equivalence	100	100	-	-	-	3.7	3.5	-0.236	.814	0.01
Internal consistency	97.4	100	1.297(1)	.255	0.12	4.1	4.4	-1.892	.059	0.04
Stability	38.5	64.0	5.734(1)	.017	0.25	3.4	3.6	-0.469	.639	< 0.01
IRT	0.00	22.2	3.234(1)	.072	0.24	-	3.9	-	-	-
Inter-rater	-	100	-		-	-	4.3	-	-	-
Scales and interpretation of scores	94.9	100	2.623	.105	0.17	3.9	4.1	-1.321	.186	0.02

Note: DIF: differential item functioning; IRT: item response theory.The average quality scores correspond to a five-point rating scale, where: 1 = Inadequate, 2 = Adequate with some deficiencies, 3 = Adequate, 4 = Good, and 5 = Excellent.The symbol (-) indicates that no information is provided or it is not applicable due to the characteristics of the test.

Conclusions

Evaluation Procedure

The test review process, led by the COP Test Commission, follows a rigorous and systematic approach that guarantees the comprehensive evaluation of the psychometric and technical properties of the measurement instruments (Fernández-Ballesteros et al., 2001). This procedure has facilitated the integration, on the one hand, of the opinion of expert reviewers in psychometrics and in the variables evaluated by the tests and, on the other hand, of the recommendations and allegations of publishers and authors, which reinforces the quality of the published reports.

It is important to highlight that, although review teams were successfully assembled for all tests, the process revealed differences in the ease of recruiting psychometricians compared to construct experts. While the psychometrician profile was filled relatively quickly, recruiting construct experts required greater effort and multiple attempts, which may reflect a lower perceived importance of the process and the relevance of the review results in their professional practice. This finding underscores the need to implement specific strategies to attract and engage these specialists, given their critical role in the comprehensive evaluation of the instruments. Although the agreement between the reviewers—the psychometrician and the content expert—of the same test varies depending on the test characteristic evaluated (Abad, 2024), both profiles contribute to an accurate and comprehensive assessment. Finally, it is essential to intensify efforts to disseminate the CET-based test review process across the various applied fields of psychology (Muñiz et al., 2020).

The results obtained in this eleventh edition show that the most recently published or updated commercial tests achieve average ratings ranging from good to excellent. In contrast, the CTS-2—a non-commercial test—and the T.A.L.E.—which has not been revised in the last forty years—record scores that indicate inadequate ratings or, at best, adequate with some shortcomings. On one hand, it is important to consider that the review of non-commercial tests poses significant challenges, such as the lack of manuals and materials, as well as the dispersion of relevant information (Abad, 2024). Moreover, the heterogeneity of the studies in terms of design, sample size and composition, and methodological quality could be contributing to the variability observed in the ratings of the psychometric properties of these tests. In relation to the T.A.L.E., considering both its lack of updates and the quality results obtained, the use of more recent instruments is recommended for evaluating literacy. An example is the *Baterías de Evaluación Cognitiva de las Dificultades en la Lectura y Escritura* [Cognitive Assessment Batteries for Reading and Writing Difficulties] (BECOLE-R; Galve Manzano & Martínez Arias, 2019), which was evaluated in the ninth edition of the test review system (Lozano, 2023), obtaining average ratings between good and excellent.

Comparative Analysis of Data Reporting and Test Quality

In the comparison between the periods before and after the implementation of the test review process, there is evidence of a significant increase in the inclusion of certain psychometric characteristics in the manuals, specifically with regard to DIF and temporal stability. However, the characteristics evaluated do not show statistically significant improvements in their quality scores.

Attributing these advances exclusively to the test evaluation process would be, at best, a risky conclusion, as multiple and varied factors could explain this trend. For example, the increase in DIF studies or in the use of IRT could reflect what has also happened at the international level. Possible factors include advances in computing and greater availability of user-friendly software that removes the need for advanced programming, as well as growing awareness of gender perspectives in research, which may have encouraged more studies on gender invariance. In addition, it is likely that test developers in Spain are more routinely incorporating CET standards, and that publishers are using CET as a model and aiming for the highest possible level, seeing it as an ideal quality standard. This approach could be motivated by the desire to obtain good ratings in reports and to present their tests as high quality products.

Be that as it may, it cannot be ruled out that this system of evaluation through the CET has incentivized the continuous updating of tests, encouraging publishers and authors to implement more advanced analyses and to strengthen the technical foundations of their instruments. These advances not only increase transparency and confidence in the quality of the tests available in Spain, but also facilitate informed decision-making by psychology professionals, thus contributing to a more rigorous and ethical practice of the profession.

Acknowledgments

We would like to express our most sincere thanks to the members of the Test Commission and the administrative staff of the COP, as well as to the individuals who reviewed the tests, to the publishing houses, and to Ismael Loinaz, corresponding author of the non-commercialized CTS2 test. Their help, collaboration, and cooperation were essential for carrying out the eleventh review of tests published in Spain.

Conflict of Interest

There is no conflict of interest.

References

Abad, F. J. (2024). Décima evaluación de test editados en España: Incorporando información sobre test no comerciales [Tenth review of tests published in Spain: Incorporating information on non-commercial tests.]. *Papeles del Psicólogo*, *45*(2), 56-64. https://www.papelesdelpsicologo.es/pdf/3033.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Psychological Association.

Arribas, D., Corral, S., & Pereña, J. (2022). *CTC-R: Cuestionario TEA Clínico - Revisado [CTC-R: Clinical ASD Questionnaire - Revised]*. Hogrefe TEA Ediciones.

Chapman, H., & 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., & 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 [Assessment of violence in intimate relationships using the Conflict Tactics Scales: factor structure and gender differences in young people]. *Psicología Conductual*, *14*, 215-233.

Departamento de I+D de Pearson Clinical & Talent Assessment [Pearson Clinical & Talent Assessment R&D Department], Hernández, A., Paradell, È., & Vallar, F. (2022). *BESS: Sistema de cribado conductual y emocional del BASC-3 [BESS: BASC-3 behavioral and emotional screening system]*. Pearson.

Echeburúa, E., Amor, P. J., Loinaz, I., & Corral, P. (2010). Escala de predicción del riesgo de violencia grave contra la pareja-revisada (EPV-R) [Severe Intimate Partner Violence Risk Prediction Scale-Revised (EPV-R)]. *Psicothema*, *22*(4), 1054-1060. https://www.psicothema.com/pdf/3840.pdf

Elosua, P., & Geisinger, K. F. (2016). Cuarta evaluación de tests editados en España: Forma y fondo [Fourth review of tests published in Spain: Form and content]. *Papeles del Psicólogo*, *37*(2), 82-88. https://www.papelesdelpsicologo.es/pdf/2693.pdf

Equipo R&D - Giunti Psychometrics España [R&D Team - Giunti Psychometrics Spain], Martínez, A., Miralles, J., & de Ancos, I. (2021). *SRP-4: Escala de psicopatía - 4a Edición [SRP-4: Psychopathy Scale - 4th Edition]*. Giunti Psychometrics.

Equipo R&D - Giunti Psychometrics España [R&D Team - Giunti Psychometrics Spain], Martínez, A. , Miralles, J. , & de Ancos, I. (2022). *MASC2: Escala de Ansiedad Multidimensional para Niños - 2a Edición [MASC2: Multidimensional Anxiety Scale for Children - 2nd Edition]*. Giunti Psychometrics.

Espejo-Navarro, A. L., & Valdivia-Ramírez, D. M. (2023). *Propiedades psicométricas de la Conflict Tactics Scale en estudiantes universitarios de Lima [Psychometric properties of the Conflict Tactics Scale in university students in Lima]*. Universidad Peruana de Ciencias Aplicadas. https://repositorioacademico.upc.edu.pe/handle/10757/658452

Evers, A., Muñiz, J., Hagemeister, C., Høstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, B. (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-Ballesteros, R., De Bruyn, E. E. J., Godoy, A., Hornke, L. F., Ter Laak, J., Vizcarro, C., ... & 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., & Muñiz, J. (2017). Quinta evaluación de tests editados en España: mirando hacia atrás, construyendo el futuro [Fifth review of tests published in Spain: Looking back, building the future]. *Papeles del Psicólogo*, *38*(3), 161-168. https://doi.org/10.23923/pap.psicol2017.2844

Gallego Rodríguez, C., & 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 [Is pornography consumption related to partner violence?: The moderating role of attitudes toward women and violence]. *Psicología Conductual*, *27*(3), 431-454.

Galve Manzano, J. L., & Martínez Arias, R. (2019). *BECOLE-R: Baterías de Evaluación Cognitiva de las Dificultades en la Lectura y Escritura. Renovado y Revisado [BECOLE-R: Cognitive Assessment Batteries for Literacy Difficulties. Revised and Updated]*. Editorial CEPE.

Gómez Sánchez, L. E. (2019). Séptima evaluación de test editados en España [Seventh review of test published in Spain]. *Papeles del Psicólogo*, *40*(3), 205-210. https://doi.org/10.23923/pap.psicol2019.2909

Graña, J. L., Andreu, J. M., Peña, M., & 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 [Factor validity and reliability of the Revised Conflict Tactics Scales (CTS2) in the Spanish adult population]. *Psicología Conductual*, *21*(3), 525-543.

Hernández, A., Elosua, P., Fernández-Hermida, J. R., & Muñiz, J. (2022). Comisión de Test: Veinticinco años velando por la calidad de los test [Test Commission: twenty-five years working on test quality]. *Papeles del Psicólogo*, *43*(1), 55-62. https://doi.org/10.23923/pap.psicol.2978

Hernández, A., Hidalgo, M. D., Hambleton, R. K., & 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., Muñiz, J., Prieto, G., & Elosua, P. (2016). Revisión del modelo para evaluar la calidad de los tests utilizados en España [Assessing the quality of tests in Spain: revision of the Spanish test review model]. *Papeles del Psicólogo*, *37*(1), 192-197. https://www.papelesdelpsicologo.es/pdf/2775.pdf

Hernández, A. , Tomás, I., Ferreres, A., & Lloret, S. (2015). Tercera evaluación de tests editados en España [Third evaluation of tests published in Spain]. *Papeles del Psicólogo*, *36*(1), 1-8. https://www.papelesdelpsicologo.es/pdf/2484.pdf

Hidalgo, M. D., & Hernández, A. (2019). Sexta evaluación de test editados en España: resultados e impacto del modelo en docentes y editoriales [Sixth test review of tests published in Spain results and impact of the model on lecturers and publishers]. *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 & 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 [A theoretical and empirical approach to the study of types of partner abusers. Descriptive analysis, assessment variables, and instruments at the Brians-2 prison]*. Ministerio del Interior, Secretaría General Técnica.

Loinaz, I., Echeburúa, E., Ortiz-Tallo, M., & Amor, P. J. (2012). Propiedades psicométricas de la Conflict Tactics Scales (CTS-2) en una muestra española de agresores de pareja [Psychometric properties of the Conflict Tactics Scales (CTS-2) in a Spanish sample of partner-violent men]. *Psicothema*, *24*(1), 142-148. https://www.psicothema.com/pi?pii=3991

Lozano, L. M. (2023). Novena evaluación de los test editados en España [Ninth review of tests published in Spain]. *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., & 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 [Spanish version of the abbreviated Yesavage questionnaire (GDS) for screening depression in people over 65: adaptation and validation]. *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., ... & Blanco, M. C. (2005). Abreviar lo breve. Aproximación a versiones ultracortas del cuestionario de Yesavage para el cribado de la depresión [Shorten what is brief. Approach to ultra-short versions of the Yesavage questionnaire for screening depression]. *Atención Primaria*, *35*(1), 14-21. https://doi.org/10.1157/13071040

Muñiz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición [International Test Commission Guidelines for test translation and adaptation: Second edition]. *Psicothema*, *25*(2), 151-157. https://doi.org/10.7334/psicothema2013.24

Muñiz, J., Fernández-Hermida, J. R., Fonseca-Pedrero, E., Campillo-Álvarez, A., & Peña-Suárez, E. (2011). Evaluación de test editados en España [Review of tests published in Spain]. *Papeles del Psicólogo*, *32*(2), 113-128. https://www.papelesdelpsicologo.es/pdf/1947.pdf

Muñiz, J., Hernández, A., & Fernández-Hermida, J. R. (2020). Utilización de los test en España: El punto de vista de los psicólogos [Test use in Spain: the psychologists' viewpoint]. *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 [Intimate partner violence and conflict resolution in LGTBI relationships in Bogotá]*. Universidad Nacional de Colombia. https://repositorio.unal.edu.co/handle/unal/69077

Redondo, N., & Graña, J. I. (2015). Consumo de alcohol, sustancias ilegales y violencia hacia la pareja en una muestra de maltratadores en tratamiento psicológico [Alcohol consumption, illegal substances, and violence toward partners in a sample of abusers undergoing psychological treatment]. *Adicciones*, *27*(1), 27-36. https://doi.org/10.20882/adicciones.191

Ponsoda, V., & Hontangas, P. (2013). Segunda evaluación de test editados en España [Second evaluation of tests published in Spain]. *Papeles del Psicólogo*, *34*(2), 82-90. https://www.papelesdelpsicologo.es/pdf/2232.pdf

Portellano, J. A., Mateos, R., Martínez-Arias, R., & Sánchez-Sánchez, F. (2021). *CUMANIN-2: Cuestionario de Madurez Neuropsicológica Infantil-2 [CUMANIN-2: Child Neuropsychological Maturity Questionnaire-2]*. Hogrefe TEA Ediciones.

Prieto, G., & Muñiz, J. (2000). Un modelo para evaluar la calidad de los test utilizados en España [A model to evaluate the quality of tests used in Spain]. *Papeles del Psicólogo*, *77*, 65-71. https://www.papelesdelpsicologo.es/resumen?pii=1102

Schittekatte, M., & 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. Oral presentation at conference: "Updates on the work of the EFPA Board of Assessment" at the 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., & 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., & Cervera, M. (1984). *T.A.L.E.: Test de análisis de lectoescritura [T.A.L.E.: Literacy assessment test]*. Machado Grupo de Distribución, S.L.

Viladrich, C., Doval, E., Penelo, E., Aliaga, J., Espelt, A., García-Rueda, R., & Angulo-Brunet, A. (2021). Octava evaluación de test editados en España: Una experiencia participativa [Eighth review of tests edited in Spain: a participative experience]. *Papeles del Psicólogo*, *42*(1), 1-9. https://doi.org/10.23923/pap.psicol2020.2937