

Article

Thirteenth Review of Tests Published in Spain

Sergio Escorial Martín 

Universidad Complutense de Madrid, Spain

ARTICLE INFO

Received: October 27, 2025

Accepted: January 14, 2026

Keywords

Test
Assessing test quality
Psychometrics
Psychometric properties
CET-R

ABSTRACT

The assessment of the quality of psychological tests is a fundamental task for guaranteeing ethical and effective professional practice. This paper presents the results of the 13th edition of the evaluation project for tests published in Spain, an ongoing initiative by the Test Commission of the General Council of the Spanish Psychological Association. In this edition, six instruments were analyzed (five commercial: ATENTO, BFQ-2, EBE, EVT-3, PPVT-5; and one non-commercial: CoPsoQ-Istas21) using the Revised Test Evaluation Questionnaire (CET-R v1.1) and an expert peer-review system. The results reveal marked heterogeneity. While the quality of the materials and theoretical foundation was generally high, critical deficiencies were noted in the reported psychometric evidence. Specifically, shortcomings were detected in some sections of the validity evidence. These findings underscore the need for professionals to critically appraise the tests they plan to use, and for publishers and authors to strengthen the psychometric justification of their instruments.

Decimotercera Evaluación de Test Editados en España

RESUMEN

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

Palabras clave

Test
Evaluación test
Psicometría
Propiedades psicométricas
CET-R

Cite this article as: Escorial, S. (2026). Thirteenth review of tests published in Spain. *Papeles del Psicólogo/Psychologist Papers*, 47(3), 150-161.

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

Correspondence: Sergio Escorial Martín sergio.escorial@psi.ucm.es 

This article is published under Creative Commons License 4.0 CC-BY-NC-ND

Tests have been described as “...the most sophisticated technology available to psychologists for the practice of their profession” (Muñiz, 2010, p. 57). Surveys published by Muñiz over the past 25 years (Muñiz & Fernández-Hermida, 2000, 2010; Muñiz et al., 2020) confirm that tests are a well-established, commonly used tool, crucial in decision-making across all areas of psychology. Their value lies in their standardized nature and their known and evaluable psychometric properties. Professionals need them because “(...) rigorous assessment is the foundation of accurate diagnoses, which in turn are key to generating effective interventions” (Muñiz, 2018, p. 23). Therefore, their psychometric properties (i.e., validity and reliability) are not a mere technicality but a fundamental pillar of effective professional practice.

It is essential for psychology professionals to understand that the mere use of a test does not guarantee good professional practice. The usefulness of these tools is directly proportional to the quality of their psychometric properties, the appropriateness of their norms, and the proposed context of use. The use of instruments with methodological deficiencies, lack of validity or reliability, or obsolete norms not only undermines the effectiveness of psychological intervention but may also pose a significant ethical risk by basing decisions critical to people’s lives on manifestly erroneous data (Prieto & Muñiz, 2000). Furthermore, a test may be excellent for one purpose but unsuitable for others (AERA, APA, & NCME, 2014).

This concern for quality and the ethical use of tests was precisely the catalyst that helped psychometrics move beyond the more theoretical aspects of the discipline toward a focus on providing solutions in applied contexts. To guide developers and users, specific frameworks have been established. The most prominent of these is undoubtedly the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). At the same time, organizations such as the International Test Commission (ITC) have also played a crucial role, by publishing specific guidelines—known as the *ITC Guidelines*—to address challenges such as the cross-cultural adaptation of tests (Hernández et al., 2020; ITC, 2018; Muñiz et al., 2013) and technology-based and digital assessment (ITC and Association of Test Publishers, 2022).

Aware of this reality, and in line with international guidelines promoted by associations such as the European Federation of Psychologists’ Associations (EFPA), the General Council of the Spanish Psychological Association (COP) assumed the responsibility for providing guidance to psychology professionals. Through its Test Commission, a systematic and periodic initiative was established over a decade ago to evaluate the quality of tests published in Spain. This project not only seeks to inform psychologists about the reliability of some of the available tools but also to promote a culture of quality among publishers and test authors, encouraging continuous improvement in the adaptation and development of new instruments (Hernández et al., 2022).

To carry out this task, the Test Commission adopted the Test Evaluation Questionnaire (*Cuestionario de Evaluación de Test*, CET, for its Spanish acronym) initially developed by Prieto and Muñiz (2000). Since then, this tool has been modified and revised to align with changes in international guidelines and to incorporate

technical advances in the test validation process, resulting in the Revised Test Evaluation Questionnaire (CET-R; Hernández et al., 2016). In subsequent years, a version of the CET-R with a number of modifications (CET-R v1.1) was published on the Test Commission’s website, as described in previous studies (Abad, 2024; Guillera & Barrios, 2025). In general terms, this instrument is based on the EFPA evaluation model and provides a standardized framework for reviewing the fundamental characteristics of a test to evaluate its overall quality. The evaluation process developed and implemented in Spain by the Test Commission is collaborative, involving not only the members of the Commission itself, but also specialist external reviewers—both experts in the test’s subject matter and experts in psychometrics—as well as the publishers themselves.

A notable development in this initiative, systematized since the 10th edition (Abad, 2024), is the inclusion of non-commercial tests. This strategic decision responds to their widespread use in professional practice (Muñiz et al., 2020) and to the fact that, as they do not have the support of a publisher, they often suffer from a lack of up-to-date psychometric reviews. The goal is to fill this information gap and offer professionals the same critical analysis that commercial tests receive. However, this process entails significant methodological challenges, as previously identified by Abad (2024), such as the frequent absence of comprehensive technical manuals, the existence of multiple versions of the same instrument, or the difficulty in delimiting the bibliographic sources necessary for the evaluation.

Since the start of this project, the results of these evaluations have been published annually, both on the Test Commission’s website (<https://www.cop.es/test/>) and in the journal *Papeles del Psicólogo*. To date, 116 tests have been evaluated. Table 1 shows the list of tests evaluated in each edition. The year of the publication and the name of each coordinator is reported, in addition to the version of the CET used for the evaluation process. As can be seen, the first four editions were conducted using the original CET by Prieto and Muñiz (2000). The following five editions used the CET-R (Hernández et al., 2016). From the 10th edition to the present, the version used has been the CET-R v1.1. The overall assessment has always been positive, and some longitudinal analyses of this process, such as the one presented in the 11th edition, even suggest that this ongoing evaluation over time may have had a tangible positive effect on the quality and quantity of psychometric information that publishers report in the manuals of the most recently published tests (Guillera & Barrios, 2025).

Following this tradition, the main objective of this article is to present the results of the 13th edition of the test evaluation. For this edition, six instruments were analyzed, covering a spectrum of domains including the comprehensive assessment of language-related aspects, the assessment of executive functions and ADHD, psychosocial risk assessment, the Big Five personality traits, and a comprehensive battery for the assessment of educational well-being. We will present the scores obtained by these six tests across the main dimensions of the quality assessment instrument used (CET-R v1.1), along with a discussion of the most notable strengths and weaknesses observed in this 13th edition.

Table 1*List of Tests Evaluated in Each Edition*

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

Note: The full name corresponding to the acronym of each test is listed in the APPENDIX.

Method

Participants

As in previous editions, the search for potential reviewers was conducted with the intention that each test be reviewed by an expert in psychometrics and by an expert in the substantive variables that each test aims to measure. To identify these experts, several strategies were combined: a) the published lists of reviewers from the 11 editions of the test evaluation published in *Papeles del Psicólogo* were consulted (Abad, 2024; Elosua & Geisinger, 2016; Fonseca & Muñiz, 2017; Gómez-Sánchez, 2019; Guillera & Barrios, 2025; Hernández et al., 2015; Hidalgo & Hernández, 2019; Lozano, 2023; Muñiz et al., 2011; Ponsoda & Hontangas, 2013; Recio-Saboya, 2026; Viladrich et al., 2021); b) data on professionals were extracted after consulting the membership lists of scientific associations (i.e., the Spanish Association for the Methodology of Behavioral Sciences [*Asociación Española de Metodología de las Ciencias del Comportamiento*, AEMCCO] and the Ibero-American Association for Research on Individual Differences [*Asociación Iberoamericana para la Investigación de las Diferencias Individuales*, AIIDI]); and c) when necessary, the list of reviewers was supplemented using the network of contacts of the coordinator of the present edition.

Initially, 12 potential reviewers were contacted, two per test according to the aforementioned specialization criteria. In this first list, we sought gender parity and geographical diversity. After the initial contact with the individuals on this first list, four of them declined to participate (a 33.33% refusal rate). Of these four individuals, two belonged to the panel of experts in psychometrics and two to the panel of experts in substantive test variables. Subsequently, four additional individuals were contacted to replace those who initially declined. All individuals contacted at this second stage agreed to participate as reviewers in this edition. Table 2 presents the final panel of reviewers who participated in the 13th edition of the evaluation of tests published in Spain.

Table 2*Panel of Reviewers for the 13th Test Evaluation*

Name	Affiliation
Alicia Franco Martínez	Autonomous University of Madrid (UAM)
David Gallardo Pujol	University of Barcelona (UB)
María Dolores Hidalgo Montesinos	University of Murcia (UMU)
Manuel Ignacio Ibáñez Ribes	Jaume I University (UJI)
Lourdes Luceño Moreno	Complutense University of Madrid (UCM)
Fabia Morales Vives	Rovira i Virgili University (URV)
David Navarro González	University of Lleida (ULI)
Ricardo Olmos Albacete	Autonomous University of Madrid (UAM)
Jesús Privado Zamorano	C.E.S. Cardenal Cisneros
Patricia Recio Saboya	National University of Distance Education (UNED)
José María Ruiz Sánchez de León	Complutense University of Madrid (UCM)
Miguel Sorrel Luján	Autonomous University of Madrid (UAM)

Instrument

For this 13th edition, the Revised Test Evaluation Questionnaire (CET-R v1.1) was used, which is the version of the standardized protocol that has been in use since the 10th edition. Like the CET-R (Hernández et al., 2016), this version is structured into three main sections. The first, General Description of the Test, collects factual and descriptive information about the test, such as publishing details, the constructs it measures, the target population, the item format, administration times, and the cost of materials. The core of the evaluation is concentrated in the second section, "Assessment of Test Characteristics," which provides a detailed analysis of four psychometric pillars: 1) Rationale and Development, 2) Validity, 3) Reliability, and 4) Norms and Interpretation of Scores. Finally, the third section, Overall Assessment, provides a quantitative summary of the scores and an open-ended qualitative section to detail the strengths, weaknesses, and recommendations for professional use of the test.

Items generally use the following rating system: N/A = *Not applicable*, 0 = *No information provided in the documentation*, 1 = *Inadequate*, 2 = *Adequate, but with some deficiencies*, 3 = *Adequate*, 4 = *Good*, and 5 = *Excellent*. The “Excellent” category is accompanied by a detailed description to guide the reviewers. Additionally, for items where more objective quantification is possible (e.g., sample sizes, reliability coefficient values, etc.), specific labels are used to facilitate a more precise assessment. Finally, the general sections summarizing the overall characteristics of the test follow this same rating scale.

It is important to note that this version refines the evaluation through three key modifications: 1) a clear distinction between essential information (the absence of which results in a score of zero) and non-critical information (which does not affect scoring); 2) specific criteria for adapted tests, requiring greater weight for psychometric evidence obtained from local samples; and 3) technical guidelines for the correct assessment of the AUC (ROC curves) in diagnostic tests. All versions of the CET are available on the website of the Test Commission (<https://www.cop.es/test/>).

Procedure

The publishers (Giunti Psychometrics, Pearson Educación, and TEA-Hogrefe) proposed to the Test Commission the five commercial tests that would be included in this 13th edition of the evaluation of tests published in Spain. In addition, as in the last three editions, the Test Commission added one non-commercial instrument to the list of tests: Version 2 of the CoPsoQ-Istas21 adapted in Spain (Moncada et al., 2014a, 2014b). Table 3 presents the information on the tests evaluated in this edition.

For the five commercial tests, once the final panel of reviewers had been established, the Spanish Psychological Association (*Colegio Oficial de Psicólogos*, COP) was responsible for

coordinating the distribution of materials to each reviewer. This distribution took place during September and October 2024. At the same time, the coordinator emailed each reviewer the CET-R v1.1, setting a four-month deadline for completion. All of the reviewers’ CET-R v1.1 questionnaires were received by mid-February 2025. Between March and May 2025, the coordinator prepared a preliminary report for each test, incorporating the assessments of both reviewers. It should be noted that no major discrepancies were observed between the two reviewers’ ratings for the same test; however, when discrepancies did arise, the coordinator determined the final rating based on the reviewers’ reasoning and the information provided in the test manual.

In May 2025, the coordinator presented the preliminary results of the reports and the process to date to the COP Test Commission. In June 2025, the preliminary reports for each test were sent to the publishers, who were given one month to submit comments or clarifications. Finally, the coordinator prepared the final version of the report, taking into account the information provided by the reviewers and by the publishers in their comments. The final reports were reviewed by a member of the Test Commission during July 2025, whose suggestions—primarily stylistic—were incorporated into the reports. All reports from this edition were published on the website of the Test Commission in September 2025 (<https://www.cop.es/test/>). Figure 1 illustrates the procedure described.

Specific Protocol for the Non-Commercial Scale CoPsoQ-Istas21

The Copenhagen Psychosocial Questionnaire (CoPsoQ-Istas21) is a tool designed to assess the main psychosocial risk factors in work settings. Since its initial development in Denmark in the late 1990s, CoPsoQ has been adapted and validated in more than 25 languages, used in hundreds of research studies, and applied in

Table 3
List of Tests Evaluated in the 13th Edition

Acronym	Name	Original authors (year of publication)	Adaptation authors (year of publication)	Publisher
ATENTO	Cuestionario para la Evaluación de las Funciones Ejecutivas y el TDAH [Questionnaire for the Assessment of Executive Functions and ADHD]	Fernando Sánchez-Sánchez & María Solar (2024)	-	Hogrefe TEA Ediciones
BFQ-2	Big Five Questionnaire - 2	Gian Vittorio Caprara, Claudio Barbaranelli, Laura Borgogni, & Michele Vecchione (2007)	Departamento R&D de Giunti Psychometrics España [R&D Department of Giunti Psychometrics Spain] (2021)	Giunti Psychometrics
CoPsoQ-Istas21-V2	Cuestionario Psicosocial de Copenhagen Versión 2 [Copenhagen Psychosocial Questionnaire, Version 2]	Tage S. Kristensen, Harald Hannerz, Annie Hogh, & Vilhelm Borg (2005)	Salvador Moncada, Clara Llorens, Neus Moreno & Emilia Molinero (2014a)	-
EBE	Batería de Evaluación del Bienestar Educativo [Educational Well-being Assessment Battery]	Jesús de la Fuente Arias & José Manuel Martínez-Vicente (2024)	-	Giunti Psychometrics
EVT-3	Expressive Vocabulary Test, Third Edition [Test de vocabulario expresivo-3]	Kathleen T. Williams (2019)	Departamento de I+D de Pearson Clinical Assessment: [Pearson Clinical Assessment R&D Department]: Ana Hernández, Èrica Paradell, Laura Saavedra, & Frédérique Vallar (2024b)	Pearson Educación
PPVT-5	Peabody Picture Vocabulary Test, Fifth Edition [Test de vocabulario en imágenes Peabody-5]	Douglas M. Dunn (2019)	Departamento de I+D de Pearson Clinical Assessment: [Pearson Clinical Assessment R&D Department]: Ana Hernández, Èrica Paradell, Laura Saavedra, & Frédérique Vallar (2024a)	Pearson Educación

thousands of psychosocial risk assessment processes in the workplace. It has become one of the most widely used assessment tools and is cited by the WHO as a method of reference (Leka et al., 2010), highlighted as an example of good practice by the European Agency for Safety and Health at Work and the International Labour Organization (ILO, 2016), and used in reports by the European Foundation for the Improvement of Living and Working Conditions (Schütte et al., 2014).

In Spain, three versions of the CoPsoQ-Istas21 are available. The first was adapted into Spanish with the collaboration of the first author of the original Danish version (Moncada et al., 2005). The second version included substantial modifications in the number and grouping of the scales (Moncada et al., 2014b). Finally, the third version of the CoPsoQ-Istas21 was also adapted into Spanish by Moncada et al. (2021). Each version has a short form for companies with fewer than 25 employees; a medium version for organizations with more than 25 employees; and a long version for research purposes. The present evaluation focused on Version 2 (medium form), as it is the only one that currently provides both the documentary material (technical manual) and the software tool necessary for this evaluation. Both resources are available on the website of the Trade Union Institute for Work, Environment, and Health (*Instituto Sindical de Trabajo, Ambiente y Salud*, ISTAS) (<https://copsoq.istas21.net/>).

Because the CoPsoQ-Istas21 Technical Manual (Moncada et al., 2014a) lacks the psychometric data necessary for the review and instead refers to Moncada et al. (2014b), a systematic literature search protocol was implemented in Scopus and WoS. Initially, 91

studies were identified, which were reduced to 35 potential documents after an initial screening. For the final selection of the 12 key studies, an inclusion criterion was applied based on theoretical relevance for the conceptualization of the construct and the provision of empirical evidence of validity and reliability. This documentary corpus consisted of 4 articles fundamental to understanding the development of the test (Kristensen et al., 2005; Moncada et al., 2005; 2014b; and 2021) and 8 articles selected for providing specific psychometric evidence (Delgado-Fernández et al., 2021; García-Iglesias et al., 2021; García-Rodríguez et al., 2015; Louzán et al., 2024; Moncada et al., 2008; Navarro et al., 2022; Useche et al., 2019; Utzet et al., 2014).

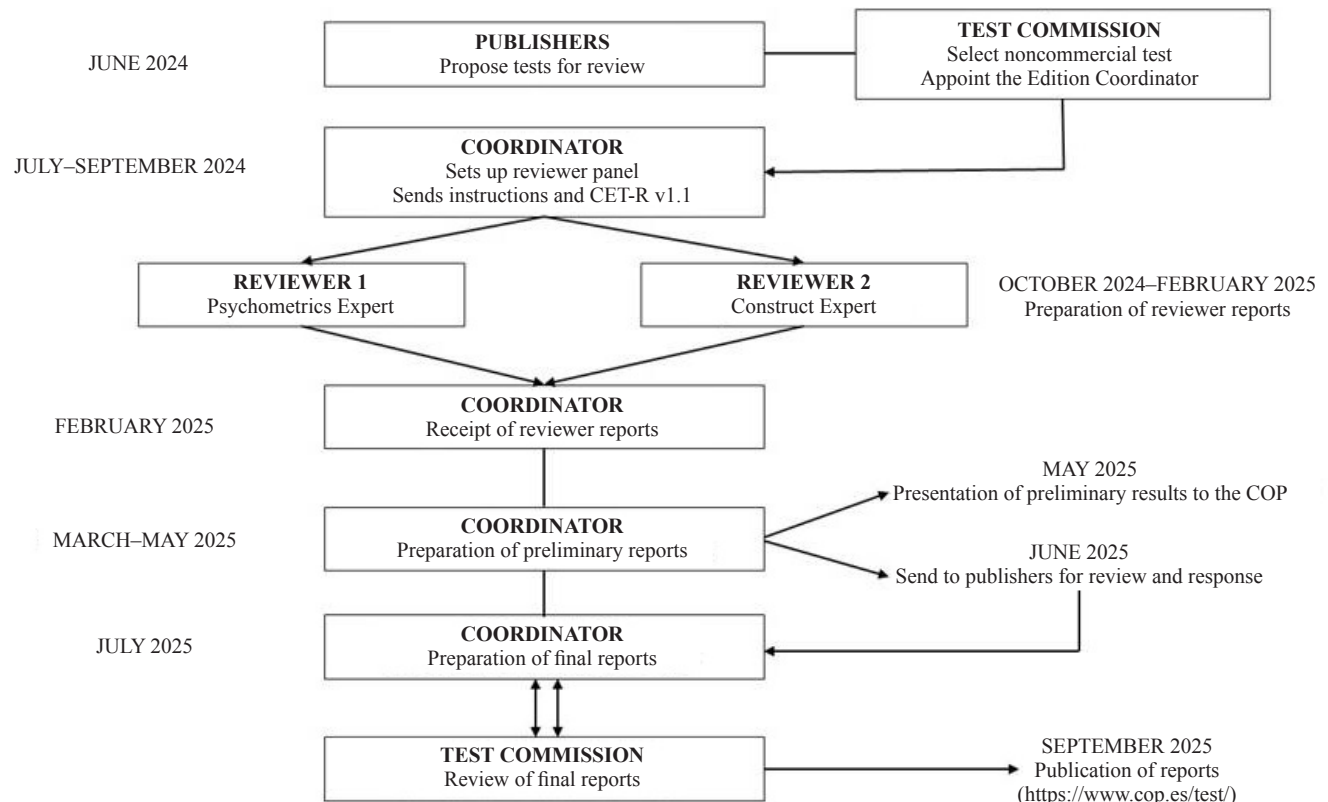
To ensure standardization of the process, reviewers were provided with a technical dossier consisting of the original materials published by ISTAS (manual and software), the full text of the 12 selected articles, and the bibliographic list of the 35 preselected studies for supplementary consultation. Once this documentation was compiled, the evaluation process was identical to the procedure followed for commercial tests (see Figure 1).

Data Analysis

First, to assess the reliability of the review process conducted in this edition, the degree of agreement among the independent reviewers was estimated using the intraclass correlation coefficient (ICC). Given the nested design of the study, in which the pairs of reviewers were independent and distinct for each instrument evaluated, a one-way random effects model was specified, analyzing

Figure 1

Representation of the Procedure for the 13th Edition of the Assessment of Tests Edited in Spain



the agreement in the individual scores. Unlike previous methodological approaches in this series of evaluations of tests published in Spain, which reported the median Pearson correlation for each evaluated test obtained from the responses of the two reviewers (i.e., [Abad, 2024](#); [Ponsoda & Hontangas, 2013](#)), in the present study we chose to perform the calculation on the dataset as a whole. This methodological decision was based on the need to avoid range-restriction bias that occurs when analyzing isolated instruments with high internal homogeneity, which tends to artificially underestimate reliability. The global analysis captures the full variance across the different levels of test quality, offering a more robust estimate of the consistency of the assessment process.

Next, the overall results obtained in the second (quantitative) section of the CET-R v1.1 were analyzed. The mean values were calculated for each of the six tests evaluated in the subsections comprising the components of this section (Rationale and Development, Validity, Reliability, and Norms and Interpretation of Scores). To contextualize the results of this edition, the average scores obtained by the six instruments evaluated in these sections (Edition Average) were compared with the accumulated historical averages. These comparisons were made at a purely descriptive level, analyzing the position of the average scores observed in this edition relative to the central tendency of two historical series: Historical 1: averages calculated from the tests of the 10th, 11th, and 12th editions, which are based on the CET-R v1.1; and Historical 2: averages calculated from the nine previous editions, the first four based on the CET and the following five on the CET-R. For each section, data were included only for tests that provided a score, which differed between the CET and the CET-R, (cumulative aggregate series). For these reasons, given the heterogeneous nature of the historical data, formal statistical inference tests were not

applied to compare mean differences. However, to facilitate a substantive interpretation, standard deviations and 95% confidence intervals were calculated for both the data from the current edition and the two historical series. In interpreting the results, differences were considered noteworthy when the mean score fell outside the confidence interval of the historical reference average, always acknowledging the descriptive nature of these assessments.

Results

The final reports for the six tests evaluated in this 13th edition are available for consultation on the website of the Test Commission, under the section for the year 2024 (<https://www.cop.es/test/>).

First, the inter-rater reliability analysis was conducted on a total of $N = 160$ pairs of valid observations, corresponding to the quantitative items of the CET-R v1.1 applicable to the six instruments. The results yielded an intraclass correlation coefficient (ICC) of 0.84, with a 95% confidence interval ranging from 0.79 to 0.88. According to the interpretation criteria established by [Cicchetti \(1994\)](#) or [Koo and Li \(2016\)](#), this value indicates an excellent level of agreement between the independent reviewers. This finding substantially exceeds the reliability estimates reported in previous editions (with medians around 0.60), suggesting that the observed variability in scores is primarily attributable to the psychometric characteristics of the assessed tests and not to the raters, thereby supporting the technical robustness of the results presented in this edition.

[Table 4](#) presents a summary of the scores obtained in each of the general sections of the second section of the CET-R v1.1. An average (Edition Average) was calculated using the scores of the six tests in each section, and two historical averages were calculated

Table 4
Scores Obtained by the Tests Evaluated in the 13th Edition

	ATENTO	BFQ-2	CoPsoQ-Istas21	EBE	EVT-3	PPVT-5	Edition Average	Historical ¹	Historical ²
Development									
Materials and documentation	5	3.5	2	4	5	5	4.1 (1.20) [2.8-5.0]	4.3 (0.67) [4.0-4.6]	4.3 (0.62) [4.2-4.4]
Theoretical foundation	5	4	4	4	3.5	3.5	4.0 (0.55) [3.4-4.6]	3.9 (1.01) [3.4-4.4]	4.1 (0.82) [4.0-4.3]
Adaptation	-	2	2	-	4.5	4.5	3.3 (1.44) [1.0-5.0]	3.9 (0.70) [3.4-4.4]	4.3 (0.70) [4.1-4.5]
Item analysis	5	3	2	3	3	3	3.2 (0.98) [2.1-4.2]	3.7 (0.88) [3.3-4.2]	3.8 (0.73) [3.7-4.0]
Validity									
Content	5	2.5	1	3	3	3	2.9 (1.28) [1.6-4.3]	3.6 (1.04) [3.1-4.1]	3.8 (0.79) [3.6-4.0]
Relationship with other variables	5	3.5	3	2.5	3.8	3.7	3.6 (0.85) [2.7-4.5]	3.5 (1.02) [3.0-4.0]	3.5 (0.64) [3.4-3.7]
Internal structure	5	2	1	2.5	0	0	1.8 (1.89) [0-3.7]	3.4 (1.12) [2.8-3.9]	3.8 (0.86) [3.6-4.0]
DIF Analysis	5	3	0	3	3	3	2.8 (1.60) [1.2-4.5]	4.1 (0.82) [3.5-4.7]	4.0 (1.02) [3.4-4.5]
Reliability									
Equivalence	-	-	-	-	-	-	-	-	3.6 (0.66) [2.9-4.3]
Internal consistency	5	4	3	3.5	4	4	3.9 (0.66) [3.2-4.6]	3.9 (0.79) [3.6-4.3]	4.2 (0.67) [4.1-4.4]
Stability	4.5	3	-	-	3.5	3.5	3.6 (0.63) [2.6-4.6]	3.3 (0.64) [3.0-3.7]	3.5 (0.78) [3.2-3.7]
IRT	-	-	-	-	-	-	-	4.0 (0.71) [3.3-4.7]	4.5 (0.45) [4.0-5.0]
Inter-rater	4	-	-	-	-	-	4.0	3.9 (1.31) [1.8-5.0]	4.5 (0.84) [3.6-5.0]
Norms and interpretation of scores	5	3	3.3	3.3	4.3	4.3	3.9 (0.78) [3.1-4.7]	3.8 (0.78) [3.5-4.2]	4.1 (0.68) [3.9-4.2]
Total	4.9	3.1	2.1	3.2	3.4	3.4			

Note: Standard deviations (SD) are shown in parentheses and 95% confidence intervals (CI) in square brackets. Comparisons with the Historical 1 column are direct, as the same version of the instrument (CET-R v1.1) was used. The descriptive statistics in the History 2 column (especially SD and CI) should be interpreted with caution as a historical reference approximation, assuming some metric heterogeneity resulting from the integration of data from the two previous versions of the instrument (CET and CET-R) in those conceptually equivalent dimensions.

to allow for purely descriptive comparisons. The first of these, corresponding to the Historical1 column was obtained from the scores of the 21 tests analyzed in the three previous editions (10th to 12th editions), since the same version of the CET-R v1.1 was used in those editions. The second, corresponding to the Historical2 column was obtained using the scores from the 89 tests analyzed in the first nine editions (first to ninth editions). The estimated standard deviation is presented in parentheses, and the 95% CI for the estimated mean is presented in square brackets. As can be seen, in some sections there are significant discrepancies with the historical pattern obtained from previous editions.

Results of the Tests Evaluated in the 13th Edition

The results obtained from the six instruments analyzed in this edition showed marked heterogeneity in the scores, both across the different dimensions of the CET-R v1.1 model and among the evaluated tests themselves (see Table 4).

With regard to the development section, the quality of the materials and documentation was the most highly rated aspect, with an overall average of 4.1. Instruments such as ATENTO, EVT-3, and PPVT-5 received the maximum score (5) in this section, standing out for the clarity and quality of their manuals. Likewise, the theoretical foundation remained at high levels (mean = 4.0), indicating that the conceptual framework of the tests was, in general, well defined. However, weaknesses were observed in some of the adaptation procedures (mean = 3.3) and in item analysis (mean = 3.2), where tests such as the CoPsoQ-Istas21 received lower scores.

The greatest discrepancies and the lowest average scores were found in the validity evidence section. While the relationship with other variables reached acceptable levels (mean = 3.6), the content validity (mean = 2.9) and, most notably, the internal structure validity (mean = 1.8) revealed significant gaps in the information provided. It is important to note that this low average for internal structure was due to the absence of updated or specific studies for several of the instruments analyzed (e.g., EVT-3 and PPVT-5 scored 0 on this specific criterion under CET-R v1.1 standards), in contrast to the excellence performance of the ATENTO test (5).

Regarding reliability, the results were generally satisfactory. Internal consistency showed robustness across most of the tests (mean = 3.9), while temporal stability (mean = 3.6) exhibited a higher CI depending on the design used to obtain this evidence for each instrument. In this reliability section, the ATENTO test stands out, not only for the high scores it obtained in the sections but also for being the only test evaluated in this edition that presented evidence of inter-rater reliability. Finally, the norms and interpretation of scores reached a good level (3.9), with most tests providing adequate and up-to-date norms. In this section, ATENTO, EVT-3, and PPVT-5 demonstrated outstanding quality. It should be noted that these tests developed their norms using continuous norming procedures that optimize the efficiency of the norming process for constructs that follow a trend across age (Evers et al., 2010; Evers et al., 2013).

When analyzing the tests as a whole, the commercial tests obtained higher total scores than the non-commercial test, CoPsoQ-Istas21, which achieved an overall score of 2.1. Notably, ATENTO achieved an overall score of 4.9, while the mean scores of the other

commercial tests were above “Adequate,” at 3.4, 3.4, 3.2, and 3.1 for EVT-3, PPVT-5, EBE, and BFQ-2, respectively.

Comparative Results of this Edition's Average With Historical Series

When comparing current results with historical averages—both the recent series under the same standard (Historical 1) and the cumulative series with previous versions of the CET (Historical 2)—divergent patterns emerge that warrant a detailed analysis. It should be noted that these comparisons are made at a descriptive level, considering the confidence intervals presented in Table 4.

In the dimensions of materials and documentation and theoretical foundation, the current edition aligns with the historical trend. The current averages are very similar to the averages in Historical 1 and Historical 2, falling within the expected margins of variability. This suggests that the formal and theoretical quality of test publishing in Spain has remained consistently high in these areas. However, in the areas of adaptation and item analysis, the averages for this edition appear to be lower, in descriptive terms, than the averages for the two historical series considered.

On the other hand, there has been a substantial decline in validity scores compared to previous editions. From a descriptive perspective, the average score for internal structure validity in this edition is lower than both Historical 1 and Historical 2. Similarly, content validity is also lower than the historical benchmarks from a descriptive standpoint. In the section on DIF analysis, a lower score was also observed descriptively for this edition. This finding does not necessarily imply a poorer intrinsic quality of the tools, but it does highlight a decrease in the comprehensiveness of the psychometric studies reported in the manuals for this edition, where several tests received scores of 0 or 1 across multiple sections (see Table 4).

Finally, in the area of reliability and norms, the data from this 13th edition are consistent with the trends of both historical series. Internal consistency reliability exactly replicates the average of the most recent series (Historical 1) and remains competitive compared to the earlier historical series (Historical 2). Similarly, the quality of the norms shows no significant differences at the descriptive level compared to the recent historical average, indicating that normative standardization continues to be a priority in test publication within our context.

Conclusions

The main objective of this study was to present the results of the 13th edition of the evaluation of tests published in Spain, continuing a project that the Test Commission of the Spanish Psychological Association (*Consejo General de la Psicología*, COP) has been carrying out for over a decade. In this edition, 6 instruments were analyzed: ATENTO, BFQ-2, CoPsoQ-Istas21, EBE, EVT-3, and PPVT-5. Their overall results display notable heterogeneity in the quality of the reported evidence. The most significant finding of this 13th evaluation is the observation of a marked dichotomy: while formal and presentation aspects maintain a good level of quality, critical deficiencies have emerged in core psychometric sections, particularly in the evidence of internal structure validity and in the DIF analyses.

On the positive side, the results of this edition are consistent with those of previous evaluations (e.g., [Abad, 2024](#); [Guillera & Barrios, 2025](#)) in showing that publishers take great care in product presentation. The average scores for Materials and Documentation (4.1) and Theoretical Foundation (4.0) clearly reflect this. Likewise, the quality of the norms (3.9) and reliability studies (3.9) also received good ratings, in line with the historical averages of previous editions. This suggests that the Spanish market offers, for the most part, well-documented tools with clear and well-founded manuals and with updated and well-developed norms—a crucial aspect for the correct interpretation of scores in accordance with international standards ([AERA, APA, & NCME, 2014](#)).

However, this positive outlook is overshadowed by the emergence of certain significant shortcomings in the psychometric evidence supporting the validity of the inferences. The average score for internal structure validity (1.8) is notably low, especially when compared to historical averages from previous editions. This data indicates that a significant proportion of the manuals evaluated in this edition did not report the factor analyses necessary to demonstrate that the test structure corresponds to the theoretical model it claims to measure. The publishers themselves also did not submit these analyses as additional evidence during the phase of review and response.

The deficiency detected in the evidence of internal structure poses a direct threat to the core validity of the test. This is an indispensable source of evidence for construct validation ([AERA, APA, & NCME, 2014](#)), an inflexible requirement that is applied with the same rigor to tests that postulate a single dimension. Unidimensionality is not an assumption but an explicit hypothesis that must be empirically demonstrated ([Brown, 2015](#)), requiring at a minimum a confirmatory factor analysis (CFA) that tests the proposed model against plausible alternatives. Assuming unidimensionality without testing it, while ignoring possible latent multidimensionality, is a methodological error: the total score becomes an uninterpretable amalgam ([Prieto & Muñiz, 2000](#)) that invalidates inferences about the trait ([Reise et al., 2013](#)). Without a solid empirical basis for internal structure, all subsequent evidence is compromised, because it is logistically impossible to know what a score correlates with if we do not know what it represents.

A similar pattern was observed for the average obtained in the differential item functioning (DIF) analysis, which in this edition was 2.8. The Standards ([AERA, APA, & NCME, 2014](#)) are explicit regarding the developers' responsibility to provide evidence of fairness and demonstrate that items do not function differently for relevant population subgroups (e.g., by gender). The lack of DIF studies poses a direct threat to the validity and fairness of assessments, an ethical issue that cannot be ignored ([Abad et al., 2011](#); [Escorial, 2017](#); [Escorial & Navas, 2007](#)).

It is essential to note that the overall averages mask significant variability, both across tests and across sections within tests. In this 13th edition, there are instruments such as ATENTO that exemplify best practices by achieving the maximum score (5) in most sections. At the opposite end of the spectrum is the non-commercial test, CoPsoQ-Istas21, with an overall average of 2.1. This disparity demonstrates that external factors such as the recency of publication or the prestige of the publishing institution do not, in themselves, constitute sufficient guarantees of the psychometric soundness of an instrument, underscoring the

necessity of a critical review of test-specific validity and reliability evidence prior to its use.

The practical implications are straightforward: a manual of high formal quality does not guarantee psychometric soundness. Practitioners have an ethical responsibility to critically analyze the evidence of validity, reliability, and norms before use. These annual evaluations do not constitute a ranking, but rather a guide to critical evaluation that highlights both strengths and, crucially, the absence of relevant information in the manuals. The usefulness of the project is validated by licensed psychologists themselves. The data indicate that those who are aware of the evaluations view them positively and consider them a necessary tool that influences their selection of instruments ([Muñiz et al., 2020](#)). However, this positive appraisal contrasts with the general lack of awareness of the project, as the same survey revealed that only 22.5% of Spanish psychologists were aware of its existence. This data highlights the need for the Test Commission to strengthen its dissemination strategies.

Furthermore, the evaluation of non-commercial tests must be maintained and strengthened. Although these public-domain instruments often exhibit greater psychometric deficiencies and receive lower ratings than commercial tests, continuing to evaluate them remains imperative. The justification is twofold: first, their frequent use in professional practice ([Muñiz et al., 2020](#)), and second, their lack of publisher support. In this regard, the fact that the low score of the CoPsoQ-Istas21 may reflect a lack of documentation and not necessarily low psychometric quality could prompt reflection on the desirability of developing a specific protocol in the future for the evaluation of non-commercial tests. The National Test Commission may debate this issue and make the decision it deems most appropriate. However, in my personal opinion, the fact that a test is free of charge does not exempt it from demonstrating quality under the same standards as the other tests used by psychology professionals in different fields ([AERA, APA, & NCME, 2014](#)). Therefore, the Test Commission is exercising its responsibility by offering this critical analysis of such tests to protect the end user.

Finally, this study has limitations. The CET-R evaluates the quality of the information reported in the manual. It does not conduct a new psychometric evaluation of the instruments. Therefore, as noted earlier, a low score may mean that the evidence does not exist or, simply, that it was not included in the documentation provided. Moreover, the project covers only a small number of tests annually and does not encompass the entire market. The instrument used (CET-R v1.1), although useful and simple, is set to be modified soon in accordance with new European guidelines. Regardless of these changes, computerizing the CET-R would bring future advantages, such as minimizing inconsistencies in scores or allowing for the weighting of items within each section.

In conclusion, the 13th edition of the test evaluation reaffirms the usefulness and relevance of this initiative led by the Test Commission. Although clear examples of excellence in test development have been observed, substantial gaps in validity evidence persist, breaking the trend observed in recent editions. Publishers and authors themselves are urged to strengthen these fundamental psychometric aspects, and professionals are urged to make critical, evidence-based choices when selecting the assessment tools that ultimately impact the well-being of the individuals they evaluate.

Acknowledgments

I would like to thank the Test Commission for its commitment to this initiative, especially Ana Hernández for her assistance during this edition; Arantxa Sánchez (COP) for her logistical support; and the publishing houses, both for providing the materials and for their constructive feedback. Finally, I express my sincere gratitude to the reviewers of this edition: their meticulous and generous work was a critical component of this project.

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.
- Abad, F. J., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en ciencias sociales y de la salud [Measurement in social and health sciences]*. Ed. Síntesis.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Psychological Association.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.
- Caprara, G. V., Barbaranelli, C., Borgogni, L., & Vecchione, M. (2007). *Big Five Questionnaire-2: Manuale*. Giunti OS: Organizzazioni Speciali.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284-290.
- de la Fuente, J., & Martínez-Vicente, J. M. (2024). *Batería para la evaluación del bienestar educativo [Battery for the evaluation of educational well-being]*. Giunti Psychometrics.
- Delgado-Fernández, V., Rey-Merchán, M. D. C., & Arquillos, A. L. (2021). Estudio comparativo de los riesgos psicosociales laborales entre profesionales médicos [Comparative study of psychosocial occupational risks among medical professionals]. *Revista de la Asociación Española de Especialistas en Medicina del Trabajo*, 30(1), 24-33.
- Departamento R&D de Giunti Psychometrics España. (2021). *BFQ-2, cuestionario "Big Five" de personalidad-2: Manual [BFQ-2, "Big Five" personality questionnaire-2: Manual] [Spanish adaptation]*. Giunti Psychometrics.
- Dunn, D. M. (2019). *Peabody Picture Vocabulary Test: PPVT-5* (5th ed.). NCS Pearson.
- 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.
- Escorial, S. (2017). Análisis de la variable sexo en la escala de Búsqueda de Sensaciones (SSS-V) empleando técnicas de Funcionamiento Diferencial de los Ítems [Analysis of the sex variable in the Sensation Seeking Scale (SSS-V) using differential item functioning techniques]. *Avances en Psicología Latinoamericana*, 35(2), 387-405.
- Escorial, S., & Navas, M. J. (2007). Analysis of the gender variable in the Eysenck Personality Questionnaire-Revised scales using differential item functioning techniques. *Educational and Psychological Measurement*, 67, 990-1001.
- Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The dutch review process for evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing*, 10, 295-317.
- Evers, A., Muñoz, J., Hagemester, C., Høstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema*, 25(3), 283-291.
- Fonseca, E., & Muñoz, 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.
- García-Iglesias, J. J., Gómez-Salgado, J., Ortega-Moreno, M., & Navarro-Abal, Y. (2021). Relationship between work engagement, psychosocial risks, and mental health among spanish nurses: A cross-sectional study. *Frontiers in Public Health*, 8, 627472.
- García-Rodríguez, A., Gutiérrez-Bedmar, M., Bellón-Saameño, J. Á., Munoz-Bravo, C., & Navajas, J. F. C. (2015). Entorno psicosocial y estrés en trabajadores sanitarios de la sanidad pública: diferencias entre atención primaria y hospitalaria [Psychosocial environment and stress in public health workers: differences between primary and hospital care]. *Atención Primaria*, 47(6), 359-366.
- Gómez-Sánchez, L. E. (2019). Séptima evaluación de test editados en España [Seventh review of tests published in Spain]. *Papeles del Psicólogo*, 40(3), 205-210.
- Guillera, G., & Barrios, M. (2025). Undécima evaluación de test editados en España [Eleventh review of tests published in Spain]. *Papeles del Psicólogo*, 46(3), 158-166.
- Hernández, A., Elosua, P., Fernández-Hermida, J. R., & Muñoz, 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.
- 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.
- Hernández, A., Paradell, È., Saavedra, L., & Vallar, F. (2024a). *Test de vocabulario en imágenes Peabody-5: PPVT-5 [Peabody Picture Vocabulary Test-5: PPVT-5] [Spanish adaptation]*. Pearson Educación.
- Hernández, A., Paradell, È., Saavedra, L., & Vallar, F. (2024b). *Test de vocabulario expresivo-3: EVT-3 [Expressive Vocabulary Test-3: EVT-3] [Spanish adaptation]*. Pearson Educación.
- Hernández, A., Ponsoda, V., Muñoz, 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.
- 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.
- Hidalgo, M. D., & Hernández, A. (2019). Sexta evaluación de test editados en España: Resultados e impacto del modelo en docentes y editoriales [Sixth review of tests published in Spain: Results and impact of the model on lecturers and publishers]. *Papeles del Psicólogo*, 40(1), 21-30.
- International Labour Organization (2016). *Workplace stress: A collective challenge*. <https://www.ilo.org/publications/workplace-stress-collective-challenge>
- International Test Commission. (2018). ITC guidelines for translating and adapting tests. *International Journal of Testing*, 18(2), 101-134.

- International Test Commission & Association of Test Publishers (2022). *Guidelines for technology based assessment*. <https://www.intestcom.org/page/28>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155-163.
- Kristensen, T. S., Hannerz, H., Høgh, A., & Borg, V. (2005). The Copenhagen Psychosocial Questionnaire—a tool for the assessment and improvement of the psychosocial work environment. *Scandinavian Journal of Work, Environment & Health*, 31(6), 438-449.
- Leka, S., Jain, A., & World Health Organization. (2010). *Health impact of psychosocial hazards at work: an overview*. World Health Organization.
- Louzán, R., & Villarroya López, A. (2024). Impact of social desirability bias on the CoPsoQ-ISTAS 21 psychosocial risk assessment questionnaire. *Work*, 79(1), 241-251.
- Lozano, L. M. (2023). Novena evaluación de los test editados en España [Ninth review of tests published in Spain]. *Papeles del Psicólogo*, 44(1), 1-7.
- Moncada, S., Llorens, C., Font, A., Galtés, A., & Navarro, A. (2008). Exposición a riesgos psicosociales entre la población asalariada en España (2004-05): valores de referencia de las 21 dimensiones del cuestionario COPSOQ ISTAS21 [Exposure to psychosocial risks among the salaried population in Spain (2004-05): Reference values of the 21 dimensions of the COPSOQ ISTAS21 questionnaire]. *Revista Española de Salud Pública*, 82, 667-675.
- Moncada, S., Llorens, C., Moreno, N., & Molinero, E. (2014a). *Manual del método CoPsoQ-istas21 (versión 2) para la evaluación y la prevención de los riesgos psicosociales en empresas con 25 o más trabajadores y trabajadoras [Manual of the CoPsoQ-istas21 method (version 2) for the assessment and prevention of psychosocial risks in companies with 25 or more workers]*. Instituto Sindical de Trabajo, Ambiente y Salud [Trade Union Institute for Work, Environment and Health].
- Moncada, S., Llorens, C., Navarro, A., & Kristensen, T. S. (2005). ISTAS21: Versión en lengua castellana del cuestionario psicosocial de Copenhague (COPSOQ) [Spanish language version of the Copenhagen Psychosocial Questionnaire (COPSOQ)]. *Archivos de Prevención de Riesgos Laborales*, 8(1), 18-29.
- Moncada, S., Llorens, C., Salas, S., Moríña, D., & Navarro, A. (2021). La tercera versión de COPSOQ-ISTAS21. Un instrumento internacional actualizado para la prevención de riesgos psicosociales en el trabajo [The third version of COPSOQ-ISTAS21. An updated international instrument for the prevention of psychosocial risks at work]. *Revista Española de Salud Pública*, 95, e1-16.
- Moncada, S., Utzet, M., Molinero, E., Llorens, C., Moreno, N., Galtés, A., & Navarro, A. (2014b). The Copenhagen Psychosocial Questionnaire II (COPSOQ II) in Spain—A tool for psychosocial risk assessment at the workplace. *American Journal of Industrial Medicine*, 57(1), 97-107.
- Muñiz, J. (2010). La teoría de los tests: Teoría clásica y teoría de respuesta a los ítems [Test theories: Classical theory and item response theory]. *Papeles del Psicólogo*, 31(1), 57-66.
- Muñiz, J. (2018). La medición de lo psicológico [The measurement of the psychological]. In Academia de Psicología de España [Spanish Academy of Psychology] (Ed.), *Psicología para un mundo sostenible Volumen II [Psychology for a sustainable world. Volume II]*. (pp. 13-29). Ed. Pirámide.
- Muñiz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los test: Segunda edición [Guidelines for the translation and adaptation of tests: Second edition]. *Psicothema*, 25(2), 151-157.
- Muñiz, J., & Fernández-Hermida, J. R. (2000). La utilización de los test en España [Test use in Spain: The psychologists' viewpoint]. *Papeles del Psicólogo*, 76, 41-49.
- Muñiz, J., & Fernández-Hermida, J. R. (2010). La opinión de los psicólogos españoles sobre el uso de los test [The opinion of Spanish psychologists on the use of tests]. *Papeles del Psicólogo*, 31(1), 108-121.
- 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.
- 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.
- Navarro, A., Fernández-Cano, M. I., Salas-Nicas, S., Llorens, C., Moríña, D., & Moncada, S. (2022). Relación entre exposición a riesgos psicosociales y salud: un estudio de cohorte mediante el COPSOQistas21 [Relationship between exposure to psychosocial risks and health: a cohort study using the COPSOQistas21]. *Gaceta Sanitaria*, 36(4), 376-379.
- 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.
- Prieto, G., & Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España [A model for evaluating the quality of tests used in Spain]. *Papeles del Psicólogo*, 77, 65-77.
- Recio-Saboya, P. (2026). Duodécima evaluación de test editados en España [Twelfth review of tests published in Spain]. *Papeles del Psicólogo*, 47(2), 106-114.
- Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, 95(2), 129-140.
- Sánchez-Sánchez, F., & Solar, M. (2024). *Cuestionario Para la Evaluación de las Funciones Ejecutivas y el TDAH: ATENTO [Questionnaire for the Evaluation of Executive Functions and ADHD: ATENTO]*. Hogrefe TEA Ediciones.
- Schütte, S., Chastang, J. F., Malard, L., Parent-Thirion, A., Vermeylen, G., & Niedhammer, I. (2014). Psychosocial working conditions and psychological well-being among employees in 34 European countries. *International Archives of Occupational and Environmental Health*, 87, 897-907.
- Useche, S. A., Montoro, L., Alonso, F., & Pastor, J. C. (2019). Psychosocial work factors, job stress and strain at the wheel: Validation of the Copenhagen Psychosocial Questionnaire (COPSOQ) in professional drivers. *Frontiers in Psychology*, 10, 1531.
- Utzet, M., Moncada, S., Molinero, E., Llorens, C., Moreno, N., & Navarro, A. (2014). The changing patterns of psychosocial exposures at work in the South of Europe: Spain as a labor market laboratory. *American Journal of Industrial Medicine*, 57(9), 1032-1042.
- Viladrich, C., Doval, E., Penelo, E., Aliaga, J., Espelt, A., García-Rueda, R., & 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.
- Williams, K. T. (2019). *Expressive Vocabulary Test: EVT-3* (3rd ed.). NCS Pearson.

Appendix. List of Full Names of the Tests Included in Table 1

Edition	Acronym	Name
2024	ATENTO	Cuestionario para la Evaluación de las Funciones Ejecutivas y el TDAH [Questionnaire for the Assessment of Executive Functions and ADHD]
	BFQ-2	Big Five Questionnaire - 2
	CoPsoQ-Istas21-V2	Cuestionario Psicosocial de Copenhague Versión 2 [Copenhagen Psychosocial Questionnaire, Version 2]
	EBE	Batería de Evaluación del Bienestar Educativo [Educational Well-being Assessment Battery]
	EVT-3	Test de Vocabulario Expresivo-3 [Expressive Vocabulary Test, Third Edition]
2023	PPVT-5	Test de Vocabulario en Imágenes Peabody-5 [Test de vocabulario en imágenes Peabody-5]
	APM	Análisis de Perfiles Motivacionales [Motivational Profile Analysis]
	TAI Matrices	Test Adaptativo de Inteligencia General (online) [Adaptive General Intelligence Test (online)]
	MACI-II	Inventario Clínico para Adolescentes Millon-II [Millon Adolescent Clinical Inventory-II]
	SBE	Screening de Bienestar Emocional para Niños y Adolescentes [Emotional Well-Being Screening for Children and Adolescents]
2022	SPMSQ	Cuestionario de Pfeiffer de Evaluación de Deterioro Cognitivo [Pfeiffer Short Portable Mental Status Questionnaire]
	TECO	Prueba para la Evaluación del Conocimiento Fonológico [Phonological Awareness Assessment Test]
	THINEME	Test de Inteligencia Emocional [Emotional Intelligence Test]
	BESS	Sistema de Cribado Conductual y Emocional [Behavioral and Emotional Screening System]
	CTC-R	Cuestionario TEA Clínico-Revisado [TEA Clinical Questionnaire-Revised]
2021	CTS-2	Versión Revisada de la Conflict Tactics Scales [Revised Conflict Tactics Scales]
	CUMANIN-2	Cuestionario de Madurez Neuropsicológica Infantil-2 [Child Neuropsychological Maturity Questionnaire-2]
	MASC2	Escala de Ansiedad Multidimensional para Niños/as [Multidimensional Anxiety Scale for Children]
	SRP 4	Escala de Psicopatía [Psychopathy Scale]
	TALE	Test de Análisis de Lectoescritura [Literacy Analysis Test]
2020	BASC-3.	Sistema de Evaluación de la Conducta de Niños y Adolescentes [Behavior Assessment System for Children and Adolescents]
	CIT	Cuestionario de Impacto del Trauma [Trauma Impact Questionnaire]
	CPF	Cuestionario de Personalidad Forense [Forensic Personality Questionnaire]
	CTE	Cuestionario de Talento Emprendedor [Entrepreneurial Talent Questionnaire]
	DABS	Escala de Diagnóstico de Conducta Adaptativa [Adaptive Behavior Diagnostic Scale]
2019	GDS	Escala de Depresión Geriátrica [Geriatric Depression Scale]
	PROLEXIA	Diagnóstico y Detección Temprana de la Dislexia [Diagnosis and Early Detection of Dyslexia]
	BAYLEY-III	Escala Bayley de Desarrollo Infantil III [Bayley Scales of Infant and Toddler Development, Third Edition]
	BECOLE-R	Batería de Evaluación Cognitiva de las Dificultades en Lectura y Escritura. Revisada y Renovada [Cognitive Assessment Battery for Reading and Writing Difficulties: Revised and Renewed]
	CAG	Cuestionario de Autoconcepto Garley [Garley Self-Concept Questionnaire]
2018	DAS	Escala de Ajuste Diádico [Dyadic Adjustment Scale]
	MacArthur	MacArthur Inventario de Desarrollo Comunicativo [MacArthur-Bates Communicative Development Inventories]
	RAVEN 2	Matrices Progresivas de Raven 2 [Raven's 2 Progressive Matrices]
	BADyG/i	Batería de Actividades Mentales Diferenciales y Generales, Nivel Infantil [Battery of Differential and General Mental Activities, Preschool Level]
	BADyG/E1-r	Batería de Actividades Mentales Diferenciales y Generales, Nivel E1 renovado [Battery of Differential and General Mental Activities, Level E1 Revised]
2017	BADyG/E2-r	Batería de Actividades Mentales Diferenciales y Generales, Nivel E2 renovado [Battery of Differential and General Mental Activities, Level E2 Revised]
	BADyG/E3-r	Batería de Actividades Mentales Diferenciales y Generales, Nivel E3 Renovado [Battery of Differential and General Mental Activities, Level E3 Revised]
	BADyG/M-r	Batería de Actividades Mentales Diferenciales y Generales, Nivel M renovado [Battery of Differential and General Mental Activities, Level M Revised]
	BADyG/S-r	Batería de Actividades Mentales Diferenciales y Generales, Nivel S renovado [Battery of Differential and General Mental Activities, Level S Revised]
	BRIEF-P	Evaluación Conductual de la Función Ejecutiva - Versión infantil [Behavior Rating Inventory of Executive Function - Preschool Version]
2016	CELF-5	Evaluación Clínica de los Fundamentos del Lenguaje, 5 [Clinical Evaluation of Language Fundamentals, Fifth Edition]
	MCMI-IV	Inventario Clínico Multiaxial de Millon, IV [Millon Clinical Multiaxial Inventory-IV]
	PECO	Prueba para la Evaluación del Conocimiento Fonológico [Phonological Awareness Assessment Test]
	TONI-4	Test de Inteligencia No Verbal, 4 [Test of Nonverbal Intelligence, Fourth Edition]
	BRIEF-2	Evaluación Conductual de la Función Ejecutiva - 2 [Behavior Rating Inventory of Executive Function, Second Edition]
2015	BYI-2	Inventarios de Beck para Niños y Adolescentes - 2 [Beck Youth Inventories, Second Edition]
	DP-3	Perfil de Desarrollo - 3 [Developmental Profile 3]
	g-R Factor	Test de Inteligencia No Verbal - Revisado [Test of Nonverbal Intelligence - Revised]
	IAES-A	Inventario de Ansiedad Escolar - Abreviado [School Anxiety Inventory - Abbreviated]
	PAIB-1-R	Prueba de Aspectos Instrumentales Básicos en Lenguaje y Matemáticas: E. Infantil (5 años), 1º y 2º E. Primaria - Renovada [Basic Instrumental Aspects Test in Language and Mathematics: Early Childhood (5 years), 1st & 2nd Grade Primary - Revised]
2014	PAIB-2-R	Prueba de Aspectos Instrumentales Básicos en Lenguaje y Matemáticas: 3º y 4º E. Primaria - Renovada [Basic Instrumental Aspects Test in Language and Mathematics: 3rd & 4th Grade Primary - Revised]
	PAIB-3-R	Prueba de Aspectos Instrumentales Básicos en Lenguaje y Matemáticas: 5º y 6º E. Primaria y 1º ESO - Renovada [Basic Instrumental Aspects Test in Language and Mathematics: 5th & 6th Grade Primary and 1st Grade Middle School - Revised]
	BADYG/E2-r	Batería de Aptitudes Diferenciales y Generales Renovado E2 [Battery of Differential and General Abilities Revised E2]
	BADYG/S-r	Batería de Aptitudes Diferenciales y Generales [Battery of Differential and General Abilities]
	BAT-7	Batería de Aptitudes de TEA [TEA Aptitude Battery]
2013	BDI-FastScreen	Inventario de Depresión de Beck FastScreen para pacientes Médicos [Beck Depression Inventory-FastScreen for Medical Patients]
	BPR	Batería de Pruebas de Razonamiento [Reasoning Test Battery]
	CESPRO	Batería para la Evaluación de las Estructuras Sintáctico-Semánticas que componen los enunciados de los problemas matemáticos y de la utilización de estrategias algorítmicas para su resolución [Battery for the Assessment of Syntactic-Semantic Structures in Mathematical Word Problems and Algorithmic Problem-Solving Strategies]
	MATRICES	MATRICES Test de Inteligencia General [MATRICES General Intelligence Test]
	MBMD	Inventario Conductual de Millon para Pacientes con Diagnóstico Médico [Millon Behavioral Medicine Diagnostic]
2012	Q-PAD	Cuestionario para la Evaluación de Problemas en Adolescentes [Adolescent Problem Evaluation Questionnaire]
	SENSORY PROFILE	Perfil Sensorial-2 [Sensory Profile 2]

Thirteenth Review of Tests

Edition	Acronym	Name
2016	BADYG-E3	Batería de Aptitudes Diferenciales y Generales Renovado E3 [Battery of Differential and General Abilities Revised E3]
	CAEPO	Cuestionario de Afrontamiento al Estrés para Pacientes Oncológicos [Stress Coping Questionnaire for Cancer Patients]
	EDI-3	Inventario de Trastornos de la Conducta Alimentaria-3 [Eating Disorder Inventory-3]
	EVAPROMES	Evaluación de los Procesos Metacognitivos en Escritura [Assessment of Metacognitive Processes in Writing]
	LAEA	Listado de Adjetivos para la Evaluación del Autoconcepto [Adjective Checklist for Self-Concept Evaluation]
	MABC-2	Batería de Evaluación del Movimiento para Niños-2 [Movement Assessment Battery for Children, Second Edition]
	MABC-2-LOC	Lista de Observación Conductual de la MABC-2 [MABC-2 Behavioral Checklist]
	NEPSY-II	Batería Neuropsicológica Infantil-II [NEPSY-II Developmental Neuropsychological Assessment]
	PAIB 2 and 3	Prueba de Aspectos Instrumentales Básicos 2 y 3 [Basic Instrumental Aspects Test 2 & 3]
	PRO 1-2 and 3	Protocolo de Evaluación del Lenguaje [Language Assessment Protocol]
	TEMT	Test de Evaluación Matemática Temprana [Early Mathematical Test]
	WISC-V	Escala de Inteligencia de Wechsler para Niños-V [Wechsler Intelligence Scale for Children, Fifth Edition]
2015	ABAS-II	Sistema de Evaluación de la Conducta Adaptativa [Adaptive Behavior Assessment System]
	BADYG/M-r	Batería de Aptitudes Diferencial y Generales. Nivel M Renovado [Battery of Differential and General Abilities, Level M Revised]
	BETA	Batería para la Evaluación de los Trastornos Afásicos [Aphasia Diagnostic Battery]
	BSI-18	Inventario Breve de 18 Síntomas [Brief Symptom Inventory 18]
	CECAD	Cuestionario Educativo-Clinico: Ansiedad y Depresión [Educational-Clinical Questionnaire: Anxiety and Depression]
	EHPAP	Evaluación de Habilidades y Potencial de Aprendizaje para Preescolares [Preschool Learning Potential and Skills Assessment]
	PAIB-1	Prueba de Aspectos Instrumentales Básicos en Lenguaje y Matemáticas [Basic Instrumental Aspects Test in Language and Mathematics]
	PECC	Prueba para la Evaluación de la Cognición Cotidiana [Everyday Cognition Assessment Test]
	SCIP	Screening del Deterioro Cognitivo en Psiquiatría [Cognitive Impairment Screening in Psychiatry]
	WMS-IV	Escala de Memoria de Wechsler -IV [Wechsler Memory Scale, Fourth Edition]
	WPPSI-IV	Escala de Inteligencia de Wechsler para Preescolar y Primaria -IV [Wechsler Preschool and Primary Scale of Intelligence, Fourth Edition]
	BCSE	Test Breve para la Evaluación del Estado Cognitivo [Brief Cognitive Status Exam]
2014	BECOLE	Batería de Evaluación Cognitiva de la Lectura y la Escritura [Cognitive Assessment Battery for Reading and Writing]
	BOHEM-3	Test Boehm de conceptos básicos [Boehm Test of Basic Concepts]
	BOHEM-3 Preschool	Test Boehm de Conceptos Básicos - 3 Preescolar [Boehm Test of Basic Concepts - 3 Preschool]
	CESQT	Cuestionario para la Evaluación del Síndrome de Quemarse por el Trabajo [Burnout Syndrome Assessment Questionnaire]
	ECLE	Evaluación de la Comprensión Lectora [Reading Comprehension Assessment]
	ESQUIZO-Q	Cuestionario Oviedo para la Evaluación de la Esquizotipia [Oviedo Schizotypy Assessment Questionnaire]
	IECI	Inventario de Estrés Cotidiano Infantil [Children's Everyday Stress Inventory]
	SOC	Escala de Dificultades de Socialización de Cantoblanco [Cantoblanco Socialization Difficulties Scale]
	TRAUMA	Test de Resistencia al Trauma [Trauma Resistance Test]
	WAIS-IV	Escala de Inteligencia de Wechsler para Adultos-IV [Wechsler Adult Intelligence Scale, Fourth Edition]
	BAI	Inventario de Ansiedad de Beck [Beck Anxiety Inventory]
	BAS-II	Escalas de Aptitudes Intelectuales [Differential Ability Scales]
2013	BDI-II	Inventario de Depresión de Beck-II [Beck Depression Inventory, Second Edition]
	CEAM	Cuestionario de Estrategias de Aprendizaje y Motivación [Learning Strategies and Motivation Questionnaire]
	CompeTEA	Cuestionario para la Evaluación de las Competencias en el Contexto Laboral [Workplace Competency Assessment Questionnaire]
	EPV-R	Escala de Predicción del Riesgo de Violencia Grave contra la Pareja - Revisada [Risk Prediction Scale for Severe Intimate Partner Violence - Revised]
	ESCOLA	Escala de Conciencia Lectora [Reading Awareness Scale]
	ESPERI	Cuestionario para la Detección de los Trastornos del Comportamiento en Niños y Adolescentes [Behavioral Disorders Screening Questionnaire for Children and Adolescents]
	Merrill-Palmer-R	Escalas de desarrollo Merrill-Palmer Revisadas [Merrill-Palmer Revised Scales of Development]
	PAI	Inventario de Evaluación de la Personalidad [Personality Assessment Inventory]
	RIAS / RIST	Escalas de Inteligencia de Reynolds / Test de Inteligencia Breve de Reynolds [Reynolds Intellectual Assessment Scales / Reynolds Intellectual Screening Test]
	WNV	Escala No Verbal de Aptitud Intelectual de Wechsler [Wechsler Nonverbal Scale of Ability]
	16PF-5	Dieciséis Factores de Personalidad, quinta edición [Sixteen Personality Factor Questionnaire, Fifth Edition]
	EFAI	Evaluación Factorial de la Aptitudes Intelectuales [Factorial Assessment of Intellectual Abilities]
2012	EVALUA	Batería Psicopedagógica [Psychopedagogical Battery]
	IGF	Batería de Inteligencia General y Factorial [General and Factorial Intelligence Battery]
	MMPI-2RF	Inventario Multifásico de la Personalidad de Minnesota-2 Reestructurado [Minnesota Multiphasic Personality Inventory-2-Restructured Form]
	NEO PI-R	Inventario de Personalidad NEO Revisado [NEO Personality Inventory-Revised]
	PROLEC-R	Batería de Evaluación de Procesos Lectores, Revisada [Assessment Battery for Reading Processes, Revised]
	WISC-IV	Escala de Inteligencia de Wechsler para Niños-IV [Wechsler Intelligence Scale for Children, Fourth Edition]