GRUPO EDITORIAL DE PSICOFUNDACIÓN

Papeles del Psicólogo (2026) 47(2) 106-114

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

Twelfth Review of Tests Published in Spain

Patricia Recio-Saboya

Universidad Nacional de Educación a Distancia (UNED), Spain

ARTICLE INFO

Received: October 24, 2025
Accepted: February 9, 2026

Keywords
Test evaluation
Psychometric properties
CET-R

ABSTRACT

Since 2010, the Test Commission of the General Council of the Spanish Psychological Association has conducted an annual evaluation of the tests published in Spain, with the aim of providing rigorous and accessible information regarding their quality. The purpose of this study is to present the results of the twelfth review of tests published in Spain. In this edition, six tests from four different publishers and one non-commercial test were assessed. These instruments were designed to measure variables such as motivation, general intelligence, cognitive impairment, emotional well-being, adolescent personality, phonological awareness, and children's emotional intelligence. As in previous editions, a systematic peer review process was followed, employing the *Test Evaluation Questionnaire - Revised* (CET-R v1.1). The evaluations were integrated into a provisional report and, after considering the claims of the publishers, a final report was prepared. The results are consistent with those of previous rounds, showing that, overall, the quality of the evaluated tests is adequate, with particular strengths in the quality of materials and documentation, the evidence of content validity and internal consistency, and the quality of the norms. Looking ahead, it would be desirable to improve consistency among reviewers and to promote ongoing training in psychometrics.

Duodécima Evaluación de Test Editados en España

Palabras clave
Evaluación de test
Propiedades psicométricas
CET-R

RESUMEN

Desde 2010, la Comisión de test del Consejo General de la Psicología realiza anualmente una evaluación de los test publicados en España, con el propósito de ofrecer información rigurosa y accesible acerca de su calidad. El objetivo de este trabajo es presentar los resultados de la Duodécima Evaluación de Test Editados en España. En esta edición se evaluaron seis test de cuatro editoriales y un test no comercial, diseñados para medir variables como motivación, inteligencia general, deterioro cognitivo, bienestar emocional, personalidad en adolescentes, conciencia fonológica e inteligencia emocional en niños. Como en ediciones anteriores, se siguió un proceso sistemático de revisión por pares, utilizando el *Cuestionario para la Evaluación de Test Revisado (CET-R v1.1)*, se integraron las evaluaciones en un informe provisional y, teniendo en cuenta las alegaciones de las editoriales, se redactó el informe final. Los resultados siguen la misma línea de evaluaciones anteriores, mostrando que, en términos generales, la calidad de los test es adecuada, destacando la calidad de los materiales y documentación, las evidencias de validez de contenido y la consistencia interna, así como la calidad de los baremos. De cara al futuro sería deseable mejorar la consistencia entre revisores, y fomentar la formación continua en psicometría.

Cite this article as: Recio-Saboya, P. (2026). Twelfth review of tests published in Spain. *Papeles del Psicólogo/Psychologist Papers, 47*(2), 106-114.
https://doi.org/10.70478/pap.psicol.2026.47.12
Correspondence: Patricia Recio Saboya reciop@psi.uned.es

In various fields of psychology, assessment through testing is a common practice, with scores derived from these instruments being used for essential purposes such as diagnosis, decision-making, counseling, intervention, and personnel selection, among others.

The psychometric quality of tests is a crucial aspect in ensuring that decisions based on test results are valid, reliable, and ethical. A test lacking sufficient evidence of validity, reliability, and normative adequacy or cultural relevance can lead to misinterpretations and, consequently, to decisions that have negative effects on the individuals being assessed. Moreover, it is important to note that no psychological measurement instrument is "universally valid" for any use, at any time, and with any population, as validity depends on the context, purpose, and characteristics of the sample (Cohen & Swerdlik, 2018).

Although the primary responsibility for analyzing the psychometric quality of a test lies with the test developer, in the seventh edition of the Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014)—a reference framework for criteria regarding test development and evaluation—the responsibility of the professional is clearly emphasized when it comes to selecting a test that adequately measures the variable of interest and possesses proven metric quality. Standard 9.0 states that the test user must know the validity evidence supporting the interpretation of scores and accept the consequences of their use. Furthermore, Standard 9.1 specifies that only individuals with the necessary training, credentials, and experience should interpret and administer tests, thereby ensuring their appropriate and ethical use. Thus, the standards make it explicit that the responsibility for the appropriate and ethical use of a test does not rest solely with the test developers or publishers, but also with the psychologists who select and administer them.

In this context, having access to rigorous and up-to-date information on the quality of tests is essential for psychology professionals to make informed decisions about their use, as they will need to critically analyze the tests' psychometric quality before selecting and administering them. To address this need, a systematic test evaluation model has been developed in Spain: the Test Evaluation Questionnaire (*Cuestionario para la Evaluación de Test*, CET; Prieto & Muñiz, 2000) and its revised version (CET-R; Hernández et al., 2016), which are based on the test review model of the European Federation of Psychologists' Associations (EFPA; Evers et al., 2013).

In addition to technical evaluations, it is essential to understand the opinions of psychology professionals themselves regarding the use of tests. To this end, three national surveys have been conducted among licensed psychologists. The results of the third survey (Muñiz et al., 2020), in which 1,248 professionals participated, show that tests are routinely used in psychological practice and that psychologists do not question their utility when administered appropriately. However, the results also reveal a clear demand for greater control and regulation of the use of tests, as well as the need for more robust training in psychometrics and assessment.

It is within this context that independent and up-to-date evaluations of the technical quality of tests take on particular importance. In Spain, this initiative has been driven by the Test Commission of the General Council of the Spanish Psychological Association (COP). Since 2010, this Commission has been carrying out a systematic evaluation project of tests published in Spain, with eleven previous reviews already completed.

The main objective of this study is to present the results obtained in the twelfth national review of tests published in Spain. The ultimate goal is to assist professionals in making decisions regarding the use of tests, as this initiative aims to provide professionals with detailed information on the quality of the instruments to guide their selection and appropriate use (Prieto & Muñiz, 2000; Muñiz et al., 2011).

To date, more than a hundred tests have been evaluated, and their reports are available for consultation and free download on the website of the General Council of the Spanish Psychological Association (https://www.cop.es/test/). This same portal includes links to the articles documenting the evaluation process for each edition, published in *Papeles del Psicólogo / Psychologist Papers* since the first edition in 2010, allowing users to track the historical evolution of this project across the different coordination teams (Muñiz et al., 2011; Ponsoda & Hontangas, 2013; Hernández et al., 2015; Elosua & Geisinger, 2016; Fonseca-Pedrero & Muñiz, 2017; Hidalgo & Hernández, 2019; Gómez, 2019; Viladrich et al., 2021; Lozano, 2023; Abad, 2024; Guilera & Barrios, 2025).

Method

Tests Evaluated

In this twelfth edition, seven tests from four leading publishers—TEA, Pearson, Habilmind, and Giunti Psychometrics—were evaluated, along with a non-commercial instrument, the Pfeiffer Screening Questionnaire for Cognitive Decline (SPMSQ). These tests address a wide range of psychological variables, including motivation, general intelligence, personality traits in adolescents, emotional well-being, cognitive impairment, phonological awareness, and emotional-intelligence (see Table 1). Furthermore, they have applications in various professional fields, such as educational, clinical, forensic, occupational, and organizational psychology, as well as in research.

The *Análisis del Perfil Motivacional* questionnaire [Motivational Profile Analysis; APM in Spanish] (Valderrama et al., 2015) is a self-report instrument consisting of 80 Likert-scale items, organized into ten scales that make up what is known as the *wheel of motives* (Valderrama, 2010; 2018). These scales are grouped into five complementary pairs: affiliation-autonomy, cooperation-power, hedonism-achievement, security-exploration, and conservation-contribution. It is designed to assess the motives that influence performance and work behavior, with applications in the contexts of career counseling, personal growth, personnel selection, and talent and diversity management.

The Spanish adaptation of the Millon Adolescent Clinical Inventory-II (MACI-II) (Hernández et al., 2023) is a self-report instrument designed to assess personality traits and clinical syndromes in adolescents aged 13 to 18. The test integrates three types of information: personality traits (stable patterns), subjective concerns (psychological distress), and clinical syndromes (acute or transient in nature). It consists of 160 true/false items and takes approximately 25 minutes to complete. Its uses include planning

Table 1
List of Measurement Instruments Analyzed in the Twelfth Test Review

Acronym	Name	Publisher	Variable of interest	Area of application	Support
APM	*Análisis de Perfiles Motivacionales* [Motivational Profile Analysis]	TEA	Motivation	Educational Psychology, Work and Organizational Psychology	Paper-and-pencil and computerized
Matrices-TAI	*Test Adaptativo de Inteligencia General* [Adaptive General Intelligence Test] (online)	TEA	General Intelligence	Clinical Psychology, Forensic Psychology, Work and Organizational Psychology	Computerized
MACI-II	*Inventario Clínico para adolescentes Millon-II* [Millon-II Clinical Inventory for Adolescents]	Pearson	Personality traits and clinical syndromes in adolescents	Clinical psychology, Educational Psychology, Forensic Psychology, Research	Paper-and-pencil and computerized
SBE	*Screening de bienestar emocional para niños y adolescentes* [Emotional well-being screening for children and adolescents]	Habilmind	Emotional well-being	Educational Psychology	Computerized
SPMSQ	*Cuestionario de Pfeiffer de evaluación de deterioro cognitivo* [Pfeiffer Cognitive Decline Assessment Questionnaire]	Non-commercial	Cognitive impairment	Neuropsychology, clinical scales	Oral administration
TECO	*Prueba para la evaluación del conocimiento fonológico* [Test for the assessment of phonological awareness]	Giunti Psychometrics	Phonological awareness	Educational Psychology	Paper-and-pencil / Oral administration
THINEME	*Test de Inteligencia Emocional* [Emotional Intelligence Test]	Giunti Psychometrics	Emotional Intelligence	Educational Psychology Clinical Psychology	Paper-and-pencil and computerized

psychotherapeutic treatment, neuropsychological and psychoeducational assessment, as well as research applications.

Test Adaptativo de Inteligencia General (Matrices-TAI) [Adaptive Test of General Intelligence] (Abad et al., 2020) is a computerized and adaptive version of the TEA Matrices test (Sánchez-Sánchez, Santamaría & Abad, 2015), designed to assess general intelligence through nonverbal abstract reasoning. It can be administered across a wide age range (6 to 74 years) and in clinical, forensic, organizational, and educational settings. Its format, based on incomplete graphic matrix items (3x3), minimizes the influence of language and is available in Spanish, Catalan, and English, incorporating visual and auditory aids.

Screening de Bienestar Emocional [Emotional Well-being Screening] (SBE) (Blanco et al., 2023) is a computerized questionnaire consisting of 65 Likert-type items, designed to assess emotional well-being in children and adolescents (ages 3-18). It assesses thirteen subdimensions grouped into three blocks: causal emotions (management of fear, sadness, anger, and anxiety/stress), emotional protective factors (optimism, self-esteem, motivation, and emotional expression), and behavioral manifestations (sleep disturbances, excessive screen use, aggression/disruptiveness, and social isolation). It offers versions adapted by developmental level: for ages 3 to 8, completed by parents or caregivers, and for ages 9 to 18 in a self-report format.

Cuestionario de Pfeiffer para detectar la existencia de deterioro cognitivo [Pfeiffer Questionnaire for Detecting Cognitive Impairment] (SPMSQ-VE) (Martínez de la Iglesia et al., 2001) is the Spanish adaptation of the Short Portable Mental Status Questionnaire (SPMSQ) (Pfeiffer, 1975). It is a brief screening instrument for cognitive impairment in individuals over 65 years of age. The questionnaire consists of 10 open-ended items that assess orientation, memory, knowledge of daily events, and serial calculation ability. Scoring is based on the number of errors made, with a range of 0 to 10. It is simple and quick to administer by primary care health care professionals, and has been recommended by the Spanish national health system as the gold standard test for detecting dementia.

Test de Evaluación de la Conciencia Fonológica (TECO) [Phonological Awareness Assessment Test] (Ramos et al., 2023) is a revised version of the PECO (Ramos & Cuadrado, 2006) designed to assess phonological awareness during the early stages of learning to read in Spanish. It assesses syllabic and phonemic levels through six activities (22 items) involving tasks of identifying, adding, and omitting phonological units. The first two activities can be administered in a group setting, and the remaining four are administered individually. Its target population consists of students in the final year of preschool, the first cycle of elementary school, and in higher grades when reading and writing difficulties are present. Its primary use is educational, although it also has applications in psycholinguistic and psychoeducational research.

Test de Inteligencia Emocional (THINEME) [Emotional Intelligence Test] (Merchán et al., 2023) is a performance test designed to assess emotional intelligence in children (ages 8-12), following the model by Mayer and Salovey (1997). The test consists of 30 items and provides an overall score and specific scores in four domains: perception, facilitation, comprehension, and emotional regulation. Its potential applications range from educational counseling to clinical and psychoeducational settings, as well as the identification of emotional profiles in children.

Participating Reviewers

For this edition, 18 potential reviewers were contacted, of whom four declined to participate in the process for various reasons; thus,

14 reviewers ultimately participated, two for each test evaluated (see Table 2). The selection of reviewers was carried out with the understanding that each test should be evaluated by two experts: one in psychometrics and another in the specific area addressed by the instrument. Whenever possible, efforts were made to ensure that the psychometrics expert also had publications related to the variable measured by the test, and that the specialist in the area evaluated by the test had published works assessing the psychometric properties of a test.

All reviewers were affiliated with universities. Most were affiliated with the field of Behavioral Science Methodology; however, due to the scope of application of the tests, evaluators from the fields of Personality, Psychological Assessment and Treatment, Basic Psychology, and Developmental and Educational Psychology also participated.

Table 2
List of Reviewers for the Twelfth Test Review

Reviewer	Department
Jesús Mª Alvarado Izquierdo	Psychobiology and Methodology in Behavioral Sciences at the Complutense University of Madrid (UCM)
Isabel Benítez Baena	Methodology of Behavioral Sciences at the University of Granada (UGR)
Sergio Escorial Martín	Psychobiology and Methodology in Behavioral Sciences at the Complutense University of Madrid (UCM)
Silvia Fernández Rivas	Basic Psychology, Psychobiology, and Methodology of Behavioral Sciences at the University of Salamanca (USAL)
Juana Gómez Benito	Social Psychology and Quantitative Psychology at the University of Barcelona (UB)
Isabel Gómez Veiga	Developmental and Educational Psychology at the National Distance Education University (UNED)
Marcela Paz González Brignardello	Personality Psychology, Assessment, and Psychological Treatment at the National University of Distance Education (UNED)
Georgina Guilera Ferré	Social Psychology and Quantitative Psychology at the University of Barcelona (UB)
Elena Navarro González	Personality Psychology, Psychological Assessment and Treatment at the University of Granada (UGR)
Francisco José Rivera de los Santos	Experimental Psychology, University of Seville (US)
Encarnación Sarriá Sánchez	Methodology of Behavioral Sciences, National University of Distance Education (UNED)
Miguel Ángel Sorrel Luján	Social Psychology and Methodology, Autonomous University of Madrid (UAM)
Purificación Sierra García	Developmental and Educational Psychology at the National University of Distance Education (UNED)
Rodrigo Schames Kreitchmann	Methodology of Behavioral Sciences, National University of Distance Education (UNED)

Description of the Instrument Used for the Evaluation

The Revised Test Evaluation Questionnaire (CET-R; Hernández et al., 2016) was used to evaluate the tests; this questionnaire is based on the test evaluation model developed by the European Federation of Psychologists' Associations (EFPA; Evers et al., 2013).

The questionnaire consists of three main sections:

1. General description of the test: composed of 28 items aimed at obtaining basic and objective information about the instrument, including publisher details, authors, publication dates, area of application, item format, target populations, administration and scoring methods, scales used, available documentation, and material costs.
2. Assessment of test characteristics: examines in detail the theoretical basis and psychometric properties of the test. This section is organized into four subsections, each ending with an open-ended question to summarize the evidence found, justify the scores, and highlight the test's strengths and weaknesses.

    - *General characteristics*: analyzes aspects such as the quality of materials, documentation, theoretical basis, adaptation, item design and development, as well as the clarity of instructions and the literature used.
    - *Validity*: evaluates the available evidence based on content, the relationship between test scores and other variables (including studies with external criteria, analyses of differences between groups, and convergent and discriminant validity), and evidence based on the test's internal structure (e.g., factor analysis and differential item functioning).
    - *Reliability*: includes various approaches to assess the accuracy of scores, such as parallel-form equivalence, internal consistency, test-retest reliability, estimates based on item response theory, and inter-rater reliability.
    - *Norms and interpretation of scores*: evaluates the quality and recency of norms for normative interpretations and the appropriateness of the procedures used to establish cutoff points in criterion-referenced tests.

3. Overall assessment of the test: this section requests a qualitative evaluation, reporting on the main strengths, limitations, recommendations for use, and areas for improvement. Additionally, a summary table is completed with the averages obtained in the different sections, allowing for an overall quantitative assessment of the instrument.

The CET-R V1.1 is available on the website of the Spanish Psychological Association (https://www.cop.es/test/).

Evaluation Process

The evaluation process followed in this edition was similar to that of previous editions. The Test Commission, in collaboration with the participating publishers (Giunti Psychometrics, Habilmind, Pearson, and TEA), selected the tests to be evaluated. In this twelfth edition, six commercially available questionnaires were evaluated, and the Test Commission also included the evaluation of a non-commercial test, the Pfeiffer Screening Questionnaire for Cognitive Decline (SPMSQ).

Once the tests to be evaluated were selected, the Test Commission appointed the person responsible for coordinating this edition of the review. Between July and September 2023, the COP sent the tests received from the publishers to the coordinator. As the tests arrived, they were read to begin the selection of reviewers. For each test, the collaboration of two experts was sought: one in psychometrics and the other in the specific construct or area assessed by the test.

In October, the coordinator contacted the potential reviewers via an initial email informing them of the assigned test and the review instrument they were to use (CET-R v1.1). Those who accepted the collaboration subsequently received detailed instructions for the task (basically reviewing the test documentation and completing the CET-R) and the deadline, set for January 2024. The COP sent each reviewer all materials related to the evaluation of the assigned test (manual, answer sheets, charts, and booklets, among others) by mail.

In the case of the non-commercial test, the coordinator compiled and emailed the reviewers relevant articles on the development, adaptation, and/or psychometric properties of the Pfeiffer Cognitive Decline Assessment Questionnaire so that they could conduct the review. The study by Martínez et al. (2001) was considered the primary article for the review, as it is the Spanish adaptation used in our country and is a very comprehensive study of its psychometric properties. In addition, the original article by Pfeiffer (1975) was sent, along with four articles aimed at providing information on the psychometric quality of the test with a Spanish sample, accompanied by a brief explanation of why they might be relevant for the review. As supplementary material, a list of 37 articles on applications of the questionnaire in a Spanish sample was sent, in which those providing data related to the test's metric quality were highlighted. The full texts of the articles on this list were not sent because, a priori, they did not provide additional information; however, the coordinator remained available to send and discuss any of them.

During the review period, the coordinator was available to address any questions that arose during the test review process.

To prepare the interim report for each test, the evaluations conducted by the two reviewers were integrated. In the qualitative sections of the CET-R, the information was synthesized by considering the contributions of the two experts and the evaluation conducted by the coordinator. In the quantitative sections, when there were minor discrepancies in the numerical scores, the arithmetic mean was used. In cases where the differences were greater, the coordinator established the final score by evaluating the reviewers' arguments and the information available in the manual. If an obvious error was detected in one of the evaluations, the final score was obtained by taking the average between the other reviewer's rating and that of the coordinator herself.

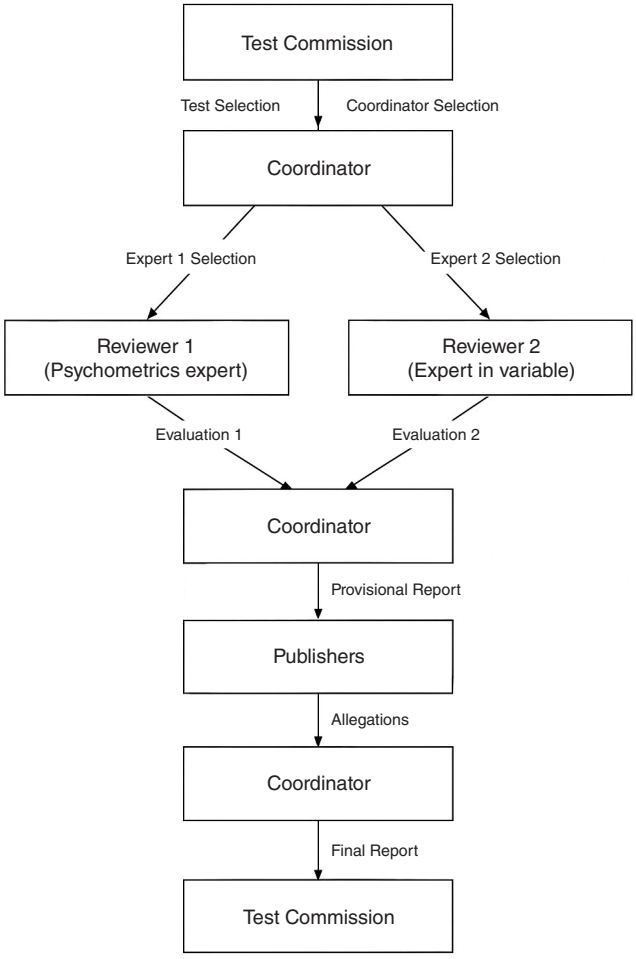
The provisional reports were sent to the COP between March and April, from where they were forwarded to the various publishers and the authors of the adaptation (in the case of the non-commercial test).

The publishers had one month to submit any comments they deemed appropriate regarding the provisional reports. In some cases, they provided additional information that was taken into account for the final report. The coordinator reviewed the comments to prepare the final report for each of the tests. Finally, a member of the COP Test Commission reviewed the final reports before they were sent for publication.

Figure 1 outlines the evaluation process.

Figure 1

Process of the Twelfth Review of Tests Published in Spain



Results

Table 3 shows the average scores obtained by the different tests in each of the dimensions analyzed. The score range varied between 1 (Inadequate) and 5 (Excellent), with two distinct options also provided for cases where no information was provided: N/A (no information was provided, but it was not considered essential given the purpose of the test) and 0 (no information was provided, yet it was considered essential given the purpose of the test).

Overall, most ratings were above 3.5, placing them in the range between adequate and good, with scores of 4 or higher being common. When comparing the results of this twelfth edition with the historical data from previous evaluations, a consistent overall trend was observed, with slight variations in some specific areas (see Table 3).

The first section covered general aspects related to the administration of the tests. In the materials and documentation section, the ratings were high, with an average of 4.0 and several tests scoring between good and excellent (Matrices-TAI = 5.0; TECO = 4.5). In theoretical basis, the results were somewhat more mixed, with scores ranging from 2.5 to 5.0, resulting in an average

Table 3
Scores Obtained on the Tests Analyzed in the Twelfth Review

Assessed characteristic	APM	Matrices-TAI	MACI-II	SBE	SPMSQ	TECO	THINEME	Average	History
Materials and documentation	4.2	5	4.5	4	3	4.5	3	4.0	4.3
Theoretical basis	4	5	4	2.5	2.5	4	4	3.7	4.1
Adaptation	---	---	4.5	---	4	---	---	4.3	4.2
Item analysis	3	5	---	3.5	3	4	3	3.6	3.8
Validity: Content	4	5	3.5	3.5	---	2.5	4	3.8	3.7
Validity: Relationship with other variables	3.1	4.5	3.5	0	3.8	3.5	2.8	3.0	3.6
Validity: Internal structure	3.8	4.8	---	4	---	2	4	3.7	3.7
Validity: DIF analysis	3.5	4.5	---	3	---	---	---	3.7	4.1
Reliability: Equivalence	---	---	---	---	---	---	---	---	---
Reliability: Internal consistency	4	5	3.5	4.5	3	4	2.8	3.8	4.2
Reliability: Stability	---	4	3	3.5	---	---	---	3.5	3.4
Reliability: IRT	---	5	---	---	---	---	---	5.0	---
Inter-rater reliability	---	---	---	---	5	---	2.5	3.8	---
Norms and interpretation of scores	4	5	4	4	3.7	3.7	3	3.9	4

Note. The scores in the table are on a scale of 0 to 5. The symbol --- indicates that no information is provided or that the item is not applicable; History is the average score from previous editions.

of 3.7, slightly lower than the historical average. Regarding adaptation, the tests that provide information achieved ratings ranging from good to excellent (MACI-II = 4.5; SPMSQ = 4.0).

Regarding validity evidence, scores for content validity were mostly adequate or good (mean = 3.8), although some heterogeneity was observed among the tests. In the evidence regarding the relationship with other variables, scores ranged from adequate to good except for two tests (one with a high rating and another with a very low one). Regarding internal structure, the mean was 3.7, in line with previous editions, although with uneven values (e.g., TECO = 2.0 vs. Matrices-TAI = 4.8). As for the DIF analysis, the ratings were positive (mean = 3.7), although this information was not always collected.

In the reliability section, the results show that internal consistency was the most widely used and highest-rated indicator, with an average of 3.8, although some tests received only adequate scores. Data on stability were more limited and showed intermediate values (mean = 3.5). In cases where analyses using the IRT (Matrices-TAI) were applied, the rating was excellent (5.0). Some ratings were also collected for inter-rater reliability, with highly variable results.

Finally, in the section on norms and interpretation of scores, the averages reached a good level (3.9), with no test evaluated below a score of 3 (adequate).

Overall, the results reflect a relatively stable pattern compared to previous evaluations, as can be seen in Table 4 and also in Figure 2, which presents the evolution of the main evaluated categories (general aspects, validity, reliability, and norms). It should be noted that this type of data must be interpreted with caution, as the number of tests to be evaluated in each edition is not very high; therefore, the average scores of evaluations in which there was a test with exceptionally high or low scores may distort the average of that evaluation. Furthermore, it should be noted that the original CET (rather than the revised version) was used in the first four evaluations, in which some headings differed—for example, there was no reliability section under item response theory.

As can be seen in Table 4, in half of the reviews there is no test that provides evidence of reliability using parallel forms, inter-rater reliability, or by providing the information function under item response theory (IRT).

Figure 2 shows that the average scores obtained in the four main sections of the CET-R, across the 12 assessments conducted to date, range from 3 (acceptable) to 4.5 (good-excellent), which generally reflects the high quality of the tests evaluated.

Discussion

The evaluation of the quality of psychological tests is an essential prerequisite for ensuring that the decisions derived from their application are valid, reliable, and ethical. The fundamental metric criteria of quality (validity and reliability) are not universal or permanent attributes of an instrument, but rather depend on the characteristics of the population, the context of use, and the purposes for which they are applied (AERA, APA & NCME, 2014). Under this premise, periodic evaluations of tests published in Spain constitute a fundamental resource, as they allow for the identification of technical strengths and limitations of the available instruments, in addition to preventing misinterpretations and protecting users from them, which reinforces trust in the tests among psychology professionals (Muñiz et al., 2020).

To date, more than a hundred tests have been evaluated, a figure that reflects the consolidation of this project since its first editions (Muñiz et al., 2011; Ponsoda & Hontangas, 2013; Hernández, Tomás, Ferreres, & Lloret, 2015), highlighting both the commitment of the Test Commission of the General Council of the Spanish Psychological Association and the interest of the scientific community in having an independent and rigorous review system.

Table 4
Average Scores Obtained by the Tests Analyzed in Each of the Reviews

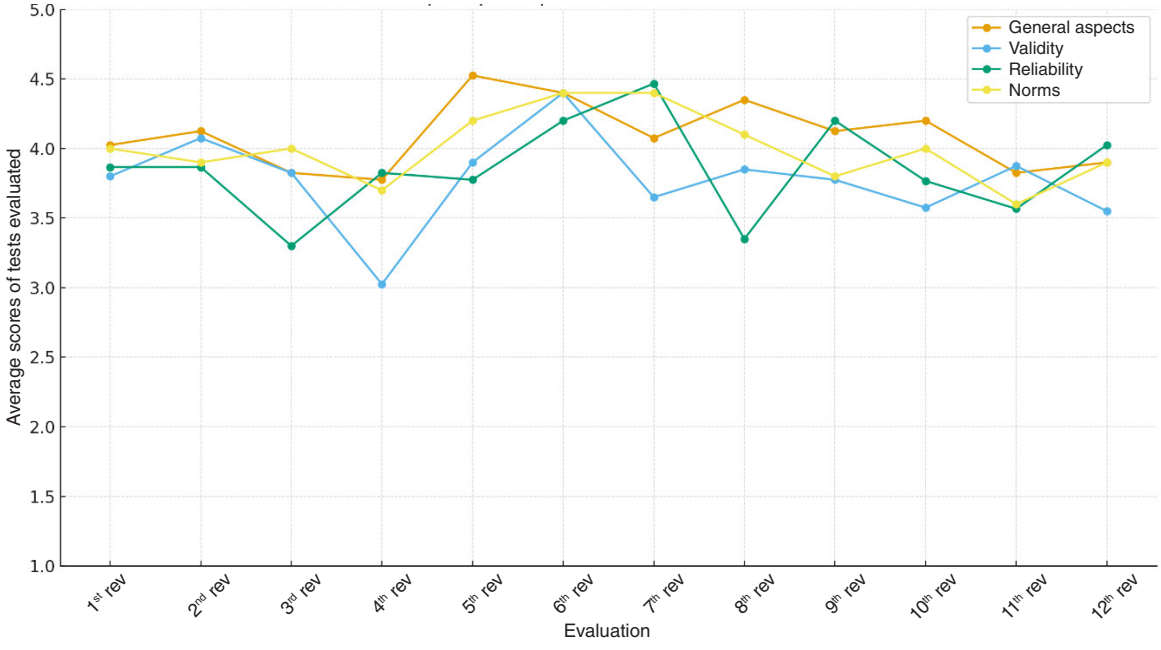
Evaluated characteristic	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th
Materials and documentation	4.3	4.3	4.1	4.0	4.5	4.6	4.1	4.3	4.5	4.5	4.5	4.0
Theoretical basis	4.1	4.0	3.6	3.6	4.7	4.6	3.9	4.7	3.9	4.2	3.7	3.7
Adaptation	4.1	4.3	3.9	3.8	4.7	4.2	4.5	4.8	4.4	4.3	3.3	4.3
Item analysis	3.6	3.9	3.7	3.7	4.2	4.2	3.8	3.6	3.7	3.8	3.8	3.6
Validity: Content	4.1	3.5	3.6	3.3	4.2	4.3	3.3	4.0	4.0	3.8	3.3	3.8
Validity: Relationship with other variables	3.6	3.6	3.6	3.2	3.6	3.8	3.7	3.3	3.6	3.7	3.8	3.0
Validity: Internal structure	3.7	4.2	3.8	3.6	3.8	4.5	3.6	3.1	4.2	3.1	3.4	3.7
Validity: DIF analysis	---	5.0	4.3	2.0	4.0	5.0	4.0	5.0	3.3	3.7	5.0	3.7
Reliability: Equivalence	4.0	4.0	---	3.0	3.0	---	4.5	3.0	---	---	---	---
Reliability: Internal consistency	3.8	4.2	3.6	4.2	4.4	4.4	4.8	4.3	4.6	4.0	4.0	3.8
Reliability: Stability	3.8	3.4	3.0	3.1	3.7	3.6	4.1	3.1	4.0	3.3	3.2	3.5
Reliability: IRT	---	---	---	---	---	4.8	---	---	4.2	4.0	3.5	5.0
Inter-rater reliability	---	---	---	---	5.0	4.0	---	---	3.0	4.0		3.8
Norms and interpretation of scores	4.0	3.9	4.0	3.7	4.2	4.4	4.4	4.1	3.8	4.0	3.6	3.9

From an applied perspective, the fact that the reports are openly available on the COP website (https://www.cop.es/test/) constitutes a valuable resource, as it allows any professional to access verified information on the psychometric quality of the instruments before deciding to use them. However, the consolidation of the project does not in itself guarantee that psychologists will take full advantage of this resource.

In this regard, the survey of psychology professionals (Muñiz et al., 2020) revealed that, although the tests are widely used and their utility is recognized, there is a clear demand for greater regulation of their use and more robust training in psychometrics. At the same time, limited knowledge of the Test Commission's work was evident, underscoring the need to intensify efforts to disseminate and transfer this resource to the professional field.

From a methodological standpoint, the use of the Test Evaluation Questionnaire (CET; Prieto & Muñiz, 2000) and its revised version (CET-R; Muñiz, Hernández & Ponsoda, 2015) in all evaluations conducted to date makes it possible to compare different editions and identify trends in the quality of tests published in Spain. However, reviewers have pointed out certain practical difficulties in applying the instrument. One of the most notable concerns the assessment of sample sizes in complex contexts, especially when the use of multiple samples raises doubts about which analyses were based on each one. Another common difficulty involves distinguishing between cases where "information is not provided, but it is not essential given the purpose of the test" and those where "information is not provided, yet it is considered essential"—a key distinction because in the first case the rating is N/A, while in the

Figure 2
Scores Obtained in Each of the Reviews Across the Four General Sections



second it must be scored as 0. At this point, reviewers with expertise in the test's content tend to experience greater uncertainty, given the psychometric knowledge required to make this decision. Likewise, discrepancies are observed among reviewers, which in some cases may be due to differences in the interpretation of the rating criteria. Given this situation, the development of a practical guide with concrete examples could facilitate the reviewers' work and increase the consistency of the evaluations.

Twelve evaluation processes have already been conducted, confirming the robustness of the test evaluation project in Spain and its value as a tool to support professional practice. The information gathered not only facilitates the informed selection of instruments but also provides publishers and authors with clear guidance on areas for improvement. Looking ahead, it would be advisable to develop practical guides to facilitate the consistent application of the CET-R in those methodological aspects that pose the greatest difficulty for reviewers, thereby contributing to improved inter-reviewer consistency. Likewise, the development of a digital version of the CET-R that allows for online administration would enable the automation of final score calculations based on partial ratings, reducing the workload and potential errors associated with the manual completion process. Finally, strengthening the dissemination of results and promoting continuing education in psychometrics and test evaluation will enable psychology professionals to fully utilize this resource and effectively integrate it into their practice, contributing to a more rigorous, responsible, and ethically sound professional practice.

Acknowledgments

First, I would like to express my gratitude to the professors from the various universities who participated as reviewers, whose work has been fundamental to the development of this evaluation. I also thank the members of the Test Commission, especially Ana Hernández for her constant support and guidance throughout the entire process. My appreciation extends to the administrative staff of the COP and the publishing houses involved. The collaboration and willingness of all of them were essential for the completion of the twelfth 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://doi.org/10.23923/pap.psicol.3033

Abad, F. J., Sánchez-Sánchez, F., & Santamaría, P. (2020). *Matrices-TAI: Test adaptativo de inteligencia general [Matrices-TAI: Adaptive General Intelligence Test]*. Hogrefe TEA Ediciones.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (7th ed.). American Educational Research Association. https://www.apa.org/science/programs/testing/standards

Blanco, I., Morón, A., Minaya, N., & Machaca-Luna, R. (2023). *Screening de Bienestar Emocional (SBE) [Emotional Well-being Screening]*. Habilmind.

Cohen, R. J., & Swerdlik, M. E. (2018). *Psychological testing and assessment: An introduction to test and measurement* (9th ed.). New York, NY: McGraw-Hill Education.

Elosua, P., & Geisinger, K. F. (2016). Cuarta evaluación de test editados en España: Forma y fondo [Fourth review of tests published in Spain: Form and content]. *Papeles del Psicólogo*, *37*(3), 82-88. https://www.papelesdelpsicologo.es/pdf/2693.pdf

Evers, A., Muñiz, J., Hagemeister, C., Hstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema*, *25*(3), 283-291. https://doi.org/10.7334/psicothema2013.97

Fonseca-Pedrero, E., & Muñiz, J. (2017). Quinta evaluación de test 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

Gómez, 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

Guilera, 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. https://www.papelesdelpsicologo.es/resumen?pii=3069

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

Hernández, A., Paradell, È., Saavedra, L., & Vallar, F. (2023). *MACI-II: Adaptación española del Millon Adolescent Clinical Inventory- II [MACI-II: Spanish adaptation of the Millon Adolescent Clinical Inventory-II]*. Pearson Educación.

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G., & Elosua, P. (2016). Revisión del modelo para evaluar la calidad de los test 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

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

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., Dueñas-Herrero, R., Onís Vilches, M. C., Aguado-Taberné, C., Colomer, C., & Luque-Luque, R. (2001). Adaptación y validación al castellano del cuestionario de Pfeiffer (SPMSQ) para detectar la existencia de deterioro cognitivo en personas mayores de 65 años [Adaptation and validation into Spanish of the Pfeiffer questionnaire (SPMSQ) to detect the existence of cognitive impairment in people over 65 years of age]. *Medicina Clínica*, *117*(4), 129-134. https://doi.org/10.1016/S0025-7753(01)72040-4

Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? In P. Salovey & D. J. Sluyter (Eds.), *Emotional development and emotional intelligence: Implications for educators* (pp. 3-31). Basic Books.

Merchán, I. M., González, J. D., & Ramos, J. L. (2023). *THINEME: Test de Habilidad en Inteligencia Emocional: Manual técnico [THINEME: Emotional Intelligence Skills Test: Technical Manual]*. Giunti Psychometrics.

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

Muñiz, J., Hernández, A., & Ponsoda, V. (2015). Quinta evaluación de test 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

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

Pfeiffer, E. (1975). A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients. *Journal of the American Geriatrics Society*, *23*(10), 433-441. https://doi.org/10.1111/j.1532-5415.1975.tb00927.x

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

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

Ramos, J. L., & Cuadrado, I. (2006). *PECO - Prueba para la Evaluación del Conocimiento Fonológico [PECO - Test for the Assessment of Phonological Awareness]*. EOS.

Ramos, J. L., González, A. I., & Gutiérrez, R. (2023). *TECO - Test de Evaluación de la Conciencia Fonológica [TECO - Test for the Assessment of Phonological Awareness]*. Giunti Psychometrics.

Sánchez-Sánchez, F., Santamaría, P., & Abad, F. J. (2015). *Matrices: Test de inteligencia general [Matrices: General Intelligence Test]*. TEA Ediciones.

Valderrama, B. (2010). *Motivación inteligente [Intelligent Motivation]*. Madrid: Prentice Hall.

Valderrama, B. (2018). La rueda de motivos: Hacia una tabla periódica de la motivación humana [The wheel of motives: Towards a periodic table of human motivation]. *Papeles del Psicólogo*, *39*(1), 60-67. https://doi.org/10.23923/pap.psicol2018.2855

Valderrama, B., Escorial, S., & Luceño, L. (2015). *APM. Análisis del Perfil Motivacional [APM: Analysis of the Motivational Profile]*. TEA Ediciones.

Viladrich, C., Doval, E., Penelo, E., Aliaga, C., Espelt, A., García-Rueda, B., & 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