GRUPO EDITORIAL DE PSICOFUNDACIÓN

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

Papeles del Psicólogo
Psychologist Papers

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

PAPELES DEL PSICÓLOGO
PSYCHOLOGIST PAPERS

Revista del Consejo General de la Psicología de España

Article

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

Francisco José Abad

Universidad Autónoma de Madrid, Spain

ARTICLE INFO

Received: January 10, 2024
Accepted: March 4, 2024

Keywords
Tests
Assessing test quality
Psychometrics
Psychometric properties
CET-R

ABSTRACT

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

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

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

RESUMEN

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

Cite this article as: Abad, F. J. (2024). Tenth review of tests published in Spain: Incorporating information on non-commercial tests. *Papeles del Psicólogo/Psychologist Papers, 45*(2), 56-64. https://doi.org/10.23923/pap.psicol.3033
Correspondence: fjose.abad@uam.es

The latest survey of Spanish psychologists on tests (Muñiz et al., 2020) shows that they are commonly used in professional practice, and they are considered to be of great help in decision making in a wide range of fields (educational, organizational, clinical, and health, etc.). Tests are recognized for their wide range of functions, such as the diagnosis, selection, orientation, and adaptation of interventions. These tools are distinguished by their standardized nature and known psychometric properties, which contribute to improved accuracy in prediction and diagnosis. However, given the vast number of tests available and the consequences that the scores can have in certain contexts, it is crucial that practitioners are able to make informed choices. Therefore, it is essential that each specific use of a test is supported by evidence.

Numerous associations and organizations have developed initiatives to ensure the quality and correct use of tests. Noteworthy among these are the "Standards for Educational and Psychological Testing", established by the main American associations of psychology and education (AERA et al., 2014). In addition, the International Test Commission (ITC) has published several guidelines to advise on key aspects of test construction and application, such as translation and adaptation (ITC, 2018; Muñiz et al., 2013), security (ITC, 2014), and the use of technologies in assessment (ITC, 2022).

Specific actions have been designed for the review and evaluation of the quality of psychological and educational tests. In the USA, the BUROS system (Carlson & Geisinger, 2012) is a notable example, while in Europe there is the test review model of the European Federation of Psychology Associations (EFPA; Evers et al., 2013), which is in the process of being updated (Schittekatte et al., 2023). In Spain, the Test Commission of the Spanish Psychological Association (COP) has been publishing reviews since 2011 (Muñiz et al., 2011). The first four editions (1st, Muñiz et al., 2011; 2nd, Ponsoda & Hontangas, 2013; 3rd, Hernández et al., 2015; 4th, Elosua & Geisenger, 2016) were carried out following an adapted version of the European model, the *Cuestionario de Evaluación de Test* [Test Evaluation Questionnaire] (CET; Prieto & Muñiz, 2000). For subsequent editions (5th, Fonseca-Pedrero & Muñiz, 2017; 6th, Hidalgo & Hernández, 2019; 7th, Gómez, 2019; 8th, Viladrich et al., 2021; 9th, Lozano, 2023), a new version of the CET questionnaire (CET-R; Hernandez et al., 2016) was used, which reflects the updates of the European model (Evers et al., 2013). To date, 96 tests have been reviewed, as detailed in Table 1. The number of different tests is somewhat lower, as some tests have required multiple assessments (e.g., for the BADYG, one report per level has been performed) or have been subject to review on more than one occasion (e.g., PAIB).

Table 1 shows that the reviews have focused mainly on commercial tests, which is consistent with the results of the survey by Muñiz et al. (2020), where most of the tests most commonly used by Spanish psychologists, except for the Geriatric Depression Scale (GDS), are commercial. To date, only one non-commercial test has been analyzed, the *Escala Revisada de Predicción del Riesgo de Violencia Grave contra la Pareja* [Revised Scale for Predicting Risk of Serious Intimate Partner Violence] (EPV-R; Echeburúa et al., 2010; reviewed in Ponsoda & Hontangas, 2013). However, many tests from academic sources are used both in and outside academia (see, for example, the *Banco de Instrumentos y Metodologías en Salud Mental* [Bank of Instruments and Methodologies in Mental Health], CIBERSAM, n.d., which lists tests such as the SDQ, applied in the *Encuesta Nacional de Salud* [National Health Survey], Ortuño et al., 2016). Therefore, the Test Commission set out to advance in the evaluation of professionally used non-commercial tests. The GDS (Yesavage & Sheikh, 1986) was evaluated in the present edition, and the *Escalas de Tácticas de Conflicto* [Conflict Tactics Scales] (CTS; Straus, 1979) is planned to be evaluated in the next edition. Note that this is not the first time that the CET-R has been used as a framework for evaluating the psychometric properties of this type of test. Beyond the evaluation of the EPV-R, by the COP, the CET-R has also been used for the revision of *Regulación Emocional* [Emotional Regulation] tests (Pérez-Sánchez et al., 2020, 2022), and, occasionally, to assess a specific psychometric property of a questionnaire (e.g., Espada et al., 2022; Reyes-Pérez et al., 2023).

Reviewing non-commercial tests involves significant challenges, such as the lack of complete manuals and the dispersion of relevant information. In the case of adaptations, clear criteria must be established to determine which versions of a test to evaluate. In

Table 1
List of Tests Evaluated by the COP in Each Edition

Year[1]	Coordinator	Test	Instruments
2024	F. J. Abad	7	**BASC-3**, CIT, CPF, CTE, DABS, **GDS**, PROLEXIA
2023	L. Lozano	6	BAYLEY-III, BECOLE-R, CAG, DAS, MacArthur, **Raven 2**
2021	C. Viladrich	11	*BADyG/*E1-r, *BADyG/E2-r*, *BADyG/E3-r*, BADyG/i, *BADyG/M-r*, *BADyG/S-r*, BRIEF-P, CELF-5, **MCMI-IV**, PECO, TONI-4
2019	L. Gómez	8	BRIEF-2, BYI-2, DP-3, Factor-g-R, IAES-A, *PAIB-1*, *PAIB-2*, *PAIB-3*
2019	M.D. Hidalgo	10	*BADyG/E2-r*, *BADyG/S-r*, BAT-7, BDI-FastScreen, BPR, CESPRO, MATRICES, MBMD, Perfil Sensorial-2, Q-PAD
2017	E. Fonseca	12	*BADYG/E3*, CAEPO, EDI-3, EVAPROMES, LAEA, MABC-2, MABC-2-LOC, NEPSY-II, *PAIB* (*2*, *3*), PRO (1-2, 3), TEMT, **WISC-V**
2016	P. Elosua	11	ABAS-II, *BADyG/M-r*, BETA, BSI-18, CECAD, EHPAP, *PAIB* (*1*), PECC, SCIP-S, WMS-IV, WPPSI-IV
2015	A. Hernández	11	BCSE, BECOLE, Boehm-3, Boehm-3 Preescolar, CESQT, ECLE, ESQUIZO-Q, IECI, SOC, TRauma, ***WAIS-IV***
2013	V. Ponsoda	12	**BAI**, BAS-II, **BDI**-II, CEAM, **CompeTEA**, ESCOLA, ESPERI, EPV-R, MPR, **PAI, RIAS, WNV**
2011	J. Muñiz	8	EFAI, EVALUA, IGF, **MMPI-2-RF**, **NEO PI-R**, **PROLEC-R**, ***WISC-IV*** , **16PF-5**

Notes. In bold type are those that, in some version, have been cited among the 25 tests most used by Spanish psychologists (Muñiz et al., 2020). In italics, the tests that have been evaluated on a second occasion or constitute a more recent version of the same battery (the most recent version is underlined); [1]Year of publication of the work describing the revision process of the corresponding edition.

selecting from the diversity of adaptations, factors such as the fidelity of the adaptation to the original instrument, widespread use in clinical practice and previous research, as well as the quality of translation and cultural adaptation should be considered. It is also important to consider the information available on the psychometric properties of the adaptation in local samples. In the case of tests originally developed in Spanish, it will be equally relevant to determine how to select the most relevant studies.

The coordinator's role in this assessment includes some additional functions, as they are the individuals responsible for selecting the appropriate documentary sources, which involves a meticulous search of academic databases and review of the relevant literature. This includes validation studies, such as factor analyses and correlations with other standardized scales, as well as norming studies, although the latter are less frequent. In addition, it is required to limit the number of papers for evaluation, including all relevant studies of the adapted version, but setting a reasonable limit in relation to the number of studies on the original questionnaire. In specific cases, it may also be useful to consult experts in the field or the test authors to access additional documentation on the psychometric properties of the test. In cases of application of the CET-R to the evaluation of non-commercial tests, there is a variety of approaches to the selection of documentation; from cases in which the documentation has been limited and proposed by the authors of the scale (case of the EPV-R, reviewed in Ponsoda & Hontangas, 2013), to cases in which the evaluation has been based on an intensive search in databases, jointly considering validation studies on the scale in different countries (e.g., Pérez-Sánchez et al., 2020).

Method

Participants

Fifteen experts were contacted for this edition, of whom one declined to participate because they had retired. Table 2 shows the 14 people who finally participated in this edition (i.e., two reviewers per test), who were selected in an attempt to maintain the criteria of gender parity and geographical diversity. In most cases, each test was reviewed by an expert with a methodological profile and another expert in the variable measured by the test. Thus, there are professors from the areas of Behavioral Sciences Methodology; Personality, Evaluation, and Psychological Treatments, and Developmental and Educational Psychology, as well as some clinical specialists. It was also taken into account that there was no conflict of interest or direct relationship with the authors.

Table 2
Reviewers Participating in the Tenth Test Review

Name	Affiliation
Juan Ramón Barrada González	University of Zaragoza
Paula Elosua Oliden	University of the Basque Country
Sergio Escorial Martín	Complutense University of Madrid
David Gallardo-Pujol	University of Barcelona
Eduardo García-Garzón	Camilo José Cela University
Ana Hernández-Baeza	University of Valencia
Alicia Eva López Martínez	University of Malaga
Estela López Nicolás	Huarte de San Juan Psychoeducational Intervention Center (Navarre)
Fabia Morales Vives	Rovira i Virgili University
Amparo Oliver Germes	University of Valencia
Mireia Orgilés Amorós	Miguel Hernández University
Patricia Recio Saboya	UNED
Francisco J. Román González	Universidad Autónoma de Madrid
Miguel Angel Sorrel Luján	Universidad Autónoma de Madrid

Instrument

The *Cuestionario para la Evaluación de los Test Revisado* [Test Evaluation Questionnaire-Revised] (CET-R; Hernández et al., 2016) was used to evaluate the tests. The CET-R consists of brief instructions in which the reviewer is offered some general and important observations, such as not considering information that is extraneous to the documentation delivered or, in the case of adapted tests, weighting differently the studies of the original version and the adapted form. On the other hand, a glossary of psychometric terms was provided to make it easier for all reviewers to assign the same meaning to them. The questionnaire consists of three main sections: General Description of the Test, Assessment of Test Characteristics, and Overall Assessment of the test.

The first section, General Description of the Test, contains 28 items that provide information on the publishing and authorship of the test (e.g., publication dates of the original test and the adaptation, manual and booklet prices), as well as information on the constructs measured, the test design (e.g., number of scales/items, response format, support, application time), and other aspects of use (e.g., areas of application, target populations, qualifications required for use).

The second section, Assessment of Test Characteristics, describes the psychometric properties of the scores and includes four sections:

a) General Evaluation of the Test: 10 items that evaluate the quality of the theoretical basis, the process of development and analysis of the items, instructions, as well as the materials and documentation provided;
b) Validity: 19 items that assess validity evidence related to content, internal structure, and the relationship of test scores with other variables, among other aspects;
c) Reliability: 14 items to assess aspects such as equivalence between parallel forms, test-retest stability, internal consistency, inter-rater reliability and, when item response theory (IRT) is applied, the information function; and
d) Norms and interpretation of scores: 9 items focused on the quality of the scoring process for normative interpretation and of the cut-off points for criterion-referenced interpretation.

The questions assess both the psychometric indices obtained (e.g., the size of the reliability coefficients) and the quality of the studies that support them. Both quantitative aspects (e.g., the number of experts in content validity studies, the sample size in internal consistency studies) and qualitative aspects (e.g., the quality of the criteria used in criterion-referenced validity or the appropriateness of the norm to the target population) are considered. Finally, three subsections (validity, reliability, norms and interpretation of scores) end with an open comments section in which the reviewer must summarize and justify the scores assigned.

Table 3
List of Tests Evaluated in the Tenth Edition

Acronym	Name	Original Author(s) (year of publication)	Author(s) adaptation (year of publication)	Publisher
BASC-3	Behavioral assessment system for children and adolescents	Reynolds & Kamphaus (2015)	Pearson Clinical & Talent Assessment R&D Department: Ana Hernández, Èrica Paradell, & Frédérique Vallar (2020)	Pearson Education
CIT	Trauma Impact Questionnaire	--	Crespo, González-Ordi, Gómez-Gutiérrez, & Santamaría (2020)	TEA Hogrefe
CPF	Forensic Personality Questionnaire	--	Medina & Sintas (2021)	Giunti Psychometrics
CTE	Entrepreneurial Talent Questionnaire	--	Valderrama (2021)	Giunti Psychometrics
DABS	Diagnostic Scale for Adaptive Behavior Diagnosis	Tassé, Schalock, Balboni, Bersani, Borthwick-Duffy, Spreat, Thissen, Widaman, & Zhang (2017)	Verdugo, Arias, & Navas (2021)	TEA Hogrefe
GDS	Geriatric Depression Scale	Yesavage & Sheikh (1986)	Martínez de la Iglesia, Onís, Dueñas, Albert, Aguado, & Luque (2002)	---
PROLEXIA	Diagnosis and Early Detection of Dyslexia	--	Cuetos, Arribas, Suárez-Coalla, & Martínez-García (2020)	TEA Hogrefe

The third section (General Assessment) contains a quantitative summary (averages) of the results of the previous section and an open section in which the strengths and weaknesses of the test, suggestions for use by professionals, and recommendations for improving the test should be reflected. In the final version of the report, published on the COP website, this assessment is presented at the beginning. In the final score assigned to each test, the quantitative items of each section are added (33): Materials and documentation (2 items), Theoretical rationale (1 item), Adaptation (1 item), Item analysis (1 item), Validity: content (2 items), Validity: relationship with other variables (9 items), Validity: internal structure (1 item), Validity: analysis of differential functioning of items (1 item), Reliability: equivalence (3 items), Reliability: internal consistency (2 items), Reliability: stability (2 items), Reliability: IRT (2 items), Reliability: inter-rater (1 item) and Scales and interpretation of scores (5 items).

The items generally adopt the following labeling system: 0 = *No information is provided in the documentation*; 1 = *Inadequate*; 2 = *Adequate, but with some deficiencies*; 3 = *Adequate*; 4 = *Good*; and 5 = *Excellent*. The 'Excellent' category includes a detailed description to guide assessors as to what this score represents in each section. In addition, for those items where more objective quantification is possible, specific labels are used to facilitate a more accurate assessment. For example, in the assessment of internal consistency, a scale is used where 3 = *Adequate* ($0.70 \leq r < 0.80$). In situations where the manual does not provide the necessary information to answer an item, some questions allow us to distinguish between cases in which the characteristic or section is not applicable to the instrument (where no score is given) and those in which, although applicable, the required information is missing (in which case a score of zero is assigned). The CET-R is available for consultation and download on the COP web page (https://www.cop.es/test/).

Procedure

As in previous reviews, the publishers (Giunti Psychometrics, Pearson Education, and TEA-Hogrefe) proposed to the COP Test Commission the tests they wanted to submit for evaluation (i.e., six tests). The selection by the publishers of the tests to be evaluated was carried out in two batches (the first four tests were sent to reviewers in July 2021 and the rest in October). In addition, the Test Committee decided to add a non-commercialized test, the GDS (Yesavage et al., 1982; Yesavage & Sheikh, 1986). The tests reviewed are shown in Table 3.

The review process for the commercial tests followed a similar protocol to that of previous editions. In each case, the coordinator contacted the reviewers and, when they accepted, the editor provided a complete copy of each test to the coordinator and to each reviewer. In addition, the coordinator sent the reviewers the CET-R, giving an extended period of four months for completion. Peer review responses to the CET-R were received up to April 2022. In March and April, the coordinator prepared an interim report for each test, integrating the assessments of both experts. There was not always agreement between the ratings, but in the case of discrepancies, the coordinator determined the final rating, taking into account the expert's reasoning and the information in the test manual. In May, the preliminary report was sent to the publishers, who had one month to submit their response. Finally, the coordinator prepared the final version of the report, taking into account the information provided by the experts and the publishers. These final reports were reviewed by a member of the COP Test Committee, whose style suggestions were incorporated before the final submission for publication (in September 2022).

Specific Protocol With Respect to the non-Commercial Scale, the GDS

With regard to the GDS, a list of the adapted versions was generated by performing a bibliographic search in the Web of Science[1] and based on the review paper by Cabañero-Martinez et al. (2007). With these criteria, a total of 103 articles and more than 20 versions were located, varying in number of items (with 15 and 30 being the most frequent lengths). Secondly, the criteria for choosing the version to review were established. The following were considered: (a) number of citations received in Google Scholar, Web of Science, and/or Scopus; (b) inclusion in review papers on adaptations of depression scales or in mental health instrument banks (e.g., CIBERSAM); (c) acknowledgement by the

1 Search terms: ("Geriatric depression scale" OR GDS* OR Yesavage) (Topic) and (Spanish OR Spain) (All Fields) and (GDS* OR Yesavage OR depression OR depresión) (Title)

original author of the scale adaptation (e.g., the version being listed on his or her web page); (d) size and representativeness of the validation samples.

Based on the information collected, the abbreviated version of Martínez de la Iglesia et al. (2002, 2005), with 15 items, seemed to be the most popular option, taking into account, for example, the number of citations of the papers, the inclusion in the review of screening instruments for depression in Spanish by Reuland et al. (2009) or the website of the author of the original version. Furthermore, this version is one of the two that were rated highest in the review by Cabañero-Martínez et al. (2007), who consider it to be the only one in which an adequate cross-cultural adaptation process is reported. Regarding the fourth criterion considered, the psychometric properties of this version were originally studied in a larger sample size than that of other adaptations (see Table 4), and a recent study provides information for obtaining normative data (Delgado-Losada et al., 2021). Another positive point to take into account is the brevity of this version, compared with the 30-item forms, which helps to reduce the problems of fatigue and inattention that commonly occur in the age group for which the scale is intended. Finally, an analysis of the item statements showed that their content was faithful to that of the original version, as opposed to other versions in which significant modifications have been made (e.g., that of Ortega-Orcos et al., 2007).

After selecting the specific version of the GDS by Martínez de la Iglesia et al. (2002, 2005), an exhaustive review of the articles citing these works was carried out using the Scopus and Web of Science (WoS) databases. A total of 224 papers were located. Finally, it was decided to select a set of 14 articles among which it is worth highlighting the three seminal articles on the development of the original version of the scale (Brink et al., 1982; Yesavage et al., 1982, 1986), as well as seven articles providing information regarding the psychometric properties of the adapted version (the most important ones: Martínez de la Iglesia et al., 2002, 2005; Lucas-Carrasco, 2012; Delgado-Losada et al., 2021), four review or synthesis papers of both the adapted version (Cabañero-Martínez et al., 2007) and the original (e.g., Balsamo et al., 2018). Additionally, the reviewers were provided with a compendium of 55 articles cited in these reviews to allow them to delve deeper into more specific aspects if they consider it necessary; these could include differential item functioning, studies on internal structure, or research on the reliability generalization of the scale.

Table 4

Different Versions of the GDS-15 and GDS-30 Scales in Spanish (Selection of Articles with More than 20 Citations in Google Scholar[5] and Samples of More Than 50 Evaluated).

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

Note. [1]Cited in Cabañero et al. (2017); [2]Cited in Reuland et al., 2009; [3]Cited in Mitchell et al., 2010; [4]Spanish adaptation cited on the original author's website: https://web.stanford.edu/~yesavage/GDS.html; [5]updated on 12/2023.

Results

The detailed reports corresponding to the tests evaluated in this tenth edition can be consulted and downloaded from the COP web page, in the section corresponding to the year 2021 (https://www.cop.es/test/#evaluados). In our case, the median of the correlation coefficients between the scores given by Reviewer 1 and Reviewer 2 in the questions in which both gave valid scores was 0.59, similar to that published by Ponsoda and Hontangas (2013) which was 0.61. This medium-low level of agreement is not atypical of this type of evaluation (Hogan et al., 2021). Some possible reasons for disagreement are presented in the Discussion section.

Table 5 shows a summary of the scores obtained in each section for each of the 7 tests. As can be seen, the pattern of results is very similar to that obtained in previous editions.

The first block of scores refers to the consideration of general and developmental aspects of the test. In the sections on materials and documentation and theoretical foundation, the commercial tests obtain scores that, in general, can be described as good or excellent (averages of 4.5 and 4.2, respectively). For the GDS, the first section could not be evaluated, due to the absence of a manual or printed booklets, but it obtained an excellent score (4.5) in the theoretical foundation section. The adaptation process was considered good or excellent in all adaptations.

The second block contains the evaluations of the evidence of validity of the tests. With regard to the evidence of content validity, the commercial tests generally had good or excellent ratings, indicating that this aspect was taken care of (e.g., they had a good theoretical foundation, through consultation with experts, item reviews, and/or pilot studies). However, quantitative and detailed information on this process was not provided in all cases, which could be desirable. A lower score was obtained for the GDS, as the reviewers consider that no such evidence was obtained.

Regarding the evidence of relationship with other variables, the scores were between adequate and excellent, and on average they were good (average = 3.7). It is relevant to highlight that several of the tests included criteria to assess the sensitivity and specificity of the proposed cut-off points, which allowed us to go beyond the normative interpretation.

Regarding the evidence related to the internal structure, the average score was below that found in previous editions (3.1 vs. 3.8). This is due to the fact that three scales received a lower score (2 = *Adequate with deficiencies*), explained by different reasons in each case, such as lack of evidence for some scales, and incomplete provision of information or unfavorable evidence for others. In the case of the GDS, it is necessary to increase the number of studies using local samples. With regard to the analysis of differential item functioning (DIF), it is worth noting that the invariance of scores across groups (e.g., age and sex) is increasingly being tested, which is essential to ensure the fairness of the assessment.

Table 5
Scores Obtained for the Tests Analyzed in the Tenth Evaluation

	BASC3	CIT	CPF	CTE	DABS	GDS	PROLEXIA	Average	Historical*
Development: Materials and documentation	4.8	5	3.5	3.8	5	--	5	4.5	4.3
Development: Theoretical foundation	4	5	2.5	3.5	5	4.5	5	4.2	4.1
Development: Adaptation	3.5	--	--	--	5	4.5	--	4.3	4.3
Development: Item analysis	--	4	2.5	3.5	4.5	3	5	3.8	3.8
Validity: content	4	5	-	3.5	5	1.5	4	3.8	3.8
Validity: relationship with other variables	3	4.8	3.5	2.8	3.7	4.3	4	3.7	3.6
Validity: internal structure	2	4.5	3	2	4.5	2	3.5	3.1	3.8
Validity: DIF analysis	--	4	--	--	3	4	--	3.7	--
Reliability: equivalence	--	--	--	--	--	--	--	--	--
Reliability: internal consistency	4.5	4.5	3	2.5	4.5	4.3	4.5	4.0	4.2
Reliability: stability	3.5	4	3.5	--	3	2.5	3.5	3.3	3.5
Reliability: IRT	--	--	3.5	--	4	4.5	--	4.0	--
Reliability: inter-rater	--	--	--	--	5	3	--	4.0	--
Scales and interpretation of scores	4	4.3	3.8	3.2	4.3	4	4.5	4.0	4.1

Note. The scores in the table follow a scale of 1 to 5: 1 = Inadequate; 2 = Adequate with deficiencies; 2.5 and above, Adequate; 3.5 and above, Good; 4.5 and above = Excellent. The symbol -- indicates that no information was provided or it was not applicable; *Average score in previous editions.

The third block collects evidence on the accuracy of the tests. As in previous editions, reliability is found to be assessed mainly by internal consistency indicators. The ratings are mostly good or excellent (the lowest value, 2.5, is adequate), which implies the use of samples of sufficient size and acceptable internal consistency values. However, it should be noted as a limitation that some of the tests do not include the internal consistency indicators for all the scales, which should be addressed in future editions of their manuals. With respect to stability, the values are somewhat lower, but can be considered adequate for all the tests (except for the GDS, which is due to the scarcity of studies with a local sample, which was penalized). For cases where IRT is applied or inter-rater reliability is calculated, good average scores are found.

The last block contains the assessments of the quality of the scales and the interpretation of the scores. On average, good scores were also obtained, but in the case of two scales, the scores were below the desired level. One important limitation is that no evidence was collected for the test's application in some of the target samples for which its use is proposed. Among the sections analyzed, the updating of the norms and the use of continuous norming stand out, which optimizes efficiency in the process of constructing the scales, especially in cases in which we work with a child and/or adolescent population and the constructs evaluated follow a trend according to age (Evers et al., 2010; Evers et al., 2013).

Conclusions

Overall Assessment and Possible Improvements

The results indicate the high quality of the tests published in Spain, with good scores and exhaustive studies of the psychometric properties, including the use of advanced techniques such as IRT or continuous norming. It is appreciated that several tests include studies on cut-off points, which enriches the interpretation of scores. However, it is recommended to increase studies on differential item functioning and to specify the hypotheses when describing the evidence of convergent and discriminant validity. It is also essential to avoid extrapolating the results of validation studies from one sample to another with different characteristics. Finally, we note that it is increasingly common to omit the norms from the manual, which makes it difficult to evaluate them. Our recommendation is that publishers provide the information as additional material for the test review process.

Evaluation of non-Commercial Tests

The evaluation of the non-commercial test involved specific challenges: (a) Some important criteria of the CET-R are not applicable to tests that do not have a manual or printed materials, which highlights the need to adapt the evaluation approach, but also leads us to recommend that researchers create this type of material; (b) Balancing the workload for reviewers with access to relevant information was complex. The approach of doing a systematic review/meta-analysis of all publications made the review unfeasible and the inclusion of only the papers on the adapted version could prove inadequate, so we opted for an intermediate position, providing as documentation the main studies on the scale's validation in local samples, while also including more general review studies. We found that reviewers tended to complement the documentation, so the work of coordination and integration was greater in this case; (c) The heterogeneity of the quality of the validation studies may be greater than in the case of commercial tests, so this is an additional factor of complexity; (d) For some criteria, the majority of studies referred to the original version, but the CET-R does not specify how much weight should be given to these studies; (e) The absence of norms in scientific publications can complicate the evaluation; in the choice of the version to be analyzed, the existence of a recent article with norms was decisive.

Regarding the CET-R and Possible Improvements to it

In our experience with the CET-R, we identified some problems, several of them mentioned in previous editions. First, not all reviewers summed the item scores to obtain the overall scores,

despite the clear instructions. This problem would be solved by providing an automatic calculation template to reviewers or by digitalizing the CET-R for online use.

Second, there was inconsistency in the handling of zero scores when information was missing in the manual. Some reviewers included them in the averages, while others did not. This discrepancy may be due to ambiguity in the CET-R instructions, which suggest averaging only the items with available information. Our review of previous evaluations showed that scores of zero were generally not assigned or considered in the averages, so we also followed this practice.

Third, agreement between the reviewers varied by criteria and was lower for some specific criteria (e.g., evidence of internal structure, quality of the norms). The discrepancies may be due to multiple reasons, some of which have been noted in previous reviews. A first reason is that content and methodological experts are sensitive to different aspects, the latter being more demanding in the application of the procedures. In other cases, discrepancies may be due to difficulties of assessment in complex cases. For example, the non-inclusion of norms in the manual was highly penalized by some reviewers, but not by others. The assessment of sample sizes can also be complex, when the test has different versions that are applied in different samples or when continuous norming is used. Finally, reviewers vary in the degree of penalization when the manual does not provide relevant information for one or more scales, when the information refers to the original versions of the scale, or when the test is proposed for use in several populations but appropriate norms are only provided for one of them. A brief guide of annotated examples would be useful in order to homogenize and facilitate assessments.

CET-R v1.1

A revised version of CET-R is now available, in which several significant improvements have been implemented. The new version distinguishes more clearly between information not presented that is essential for assessing the quality of a test and missing information that is not essential for this purpose. In addition, for adapted tests, reviewers are asked to specify the origin of the samples in the various sections (Item Analysis, Validity, etc.), thus making it possible to assess the degree of validation of the test with local samples. The need to use local samples for the norms to be adequate is also emphasized. Finally, guidelines are incorporated for the assessment of the area under the curve (AUC) in the use of ROC curves, an increasingly relevant aspect in studies of the sensitivity and specificity of a test, especially in the prediction of specific criteria, such as diagnostic categories.

Final Conclusions

In conclusion, it is beneficial to reflect on the impact of the test review process. An overall positive effect has been observed, particularly in the more detailed presentation of evidence supporting the technical quality of tests in recent manuals. Not only does the CET-R model guide authors and publishers in the development and adaptation of tests, but it also contributes to the dissemination of lesser-known but widely used tests that are supported by evidence of technical and psychometric quality. Furthermore, it is an important training tool for future psychologists, making them aware of the standards they must demand in the application of tests.

Regarding how to improve the knowledge of these processes among Spanish psychologists, it would be helpful to create a database of evaluated tests, organized by constructs or assessment areas, thus facilitating comparisons. The COP has already taken a step in this direction with its *Buscador de test* [test search engine] https://www.jornadas.cop.es/evaluacionTest/, which allows searches by key words and improves access to these assessments. Moreover, it is crucial to evaluate how psychologists use these reports and their practical usefulness, as well as to understand the criteria they use when choosing a test.

Finally, it is essential to persevere in the review of non-commercial tests that are used professionally, despite the difficulties inherent in this process. Such evaluation is key because if the review is favorable it enriches the set of tools available to professionals, while if it is unfavorable it mitigates the risks of using inadequate tests, based on obsolete standards or validated on unsuitable samples.

Acknowledgments

I am grateful for the collaboration of the members of the Test Committee, in particular Ana Hernández and Paula Elosua, for their valuable help and cooperation throughout the process. I would also like to express my gratitude to the publishers for providing the copies to be evaluated, as well as for providing detailed and constructive feedback in their responses to the interim reports. Finally, I would like to acknowledge and thank the reviewers for their generous and excellent collaboration, without which these evaluations would not be possible.

Conflict of Interest

There is no conflict of interest.

References

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

Abizanda, P., Luengo, C., López, J., Sánchez, P., Romero, L., & Fernández, C. (1998). Predictores de mortalidad, deterioro funcional e ingreso hospitalario en una muestra de ancianos residentes en la comunidad [Predictors of mortality, functional deterioration, and hospital admission in a sample of community-dwelling elderly people]. *Revista Española de Geriatría y Gerontología*, *33*, 219-225.

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

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

Cabañero-Martínez, M. J., Richart-Martínez, M., Muñoz-Mendoza, C. L., & Reig-Ferrer, A. (2007). Revisión estructurada de las escalas de depresión en personas mayores [Structured review of depression scales in elderly people]. *International Journal of Clinical and Health Psychology*, *7*(3), 823-846.

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

Centro de Investigación Biomédica en Red de Salud Mental [Center for Biomedical Research in the Mental Health Network] (CIBERSAM) (n.d.). *Banco de Instrumentos y Metodologías en Salud Mental. [Bank of Instruments and Methodologies in Mental Health]*. Ministerio de Ciencia e Innovación [Ministry of Science and Innovation]. https://bi.cibersam.es/busqueda-de-instrumentos

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

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

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

Dios, R. de, Hernández, A. M., Rexach, L. I., & Cruz, A. J. (2001). Validación de una versión de cinco ítems de la Escala de Depresión Geriátrica de Yesavage en una población española [Validation of a five-item version of the Yesavage Geriatric Depression Scale in a Spanish population]. *Revista Española de Geriatría y Gerontología*, *36*, 276-280. https://doi.org/10.1016/S0211-139X(01)74736-1

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

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/Psychologist Papers*, *37*(2), 82-88. https://www.papelesdelpsicologo.es/pdf/2693.pdf

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

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing*, *10*, 295-317. https://psycnet.apa.org/doi/10.1080/15305058.2010.518325

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

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

Fonseca-Pedrero, E., & 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/Psychologist Papers*, *37*(1), 161-168. https://doi.org/10.23923/pap.psicol2017.2844

García-Serrano, M. J., & Tobías, J. (2001). Prevalencia de depresión en mayores de 65 años. Perfil del anciano de riesgo [Prevalence of depression in people over 65 years of age. Profile of the elderly at risk]. *Atención Primaria*, *27*, 484-488. https://doi.org/10.1016/S0212-6567(01)78839-7

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

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G., & 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/Psychologist Papers*, *37*, 192-197. https://www.papelesdelpsicologo.es/pdf/2775.pdf

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

Hidalgo, M. D., & Hernández, A. (2019). Sexta evaluación de test editados en España: Resultados e impacto del modelo en docentes y editoriales [Sixth review of tests published in Spain: Results and impact of the model on lecturers and publishers]. *Papeles del Psicólogo/Psychologist Papers*, *40*(1), 21-30. https://doi.org/10.23923/pap.psicol2019.2886

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

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

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

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

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

Lozano, L. M. (2023). Novena evaluación de los test editados en España [Ninth review of tests published in Spain]. *Papeles del Psicólogo/Psychologist Papers*, *44*(1), 1-7. https://doi.org/10.23923/pap.psicol.3004

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

Martí, D., Miralles, R., Llorach, I., García-Palleiro, P., Esperanza, A., Guillén, J., & Cervera, A. (2000). Trastornos depresivos en una unidad de convalecencia: experiencia y validación de una versión española de 15 preguntas de la escala de depresión geriátrica de Yesavage [Depressive disorders in a convalescent unit: Experience and validation of a 15-question Spanish version of the Yesavage geriatric depression scale]. *Revista Española de Geriatría y Gerontología*, *35*, 1-7.

Martínez de la Iglesia, J., Onís, M. C., Dueñas, H. R., Albert, C. C., Aguado, T. C., & Luque, L. R. (2002). Versión española del cuestionario de Yesavage abreviado (GDS) para el despistaje de depresión en mayores de 65 años: adaptación y validación [Spanish version of the abbreviated Yesavage questionnaire (GDS) for depression screening in people over 65 years of age: Adaptation and validation]. *Revista de Medicina Familiar y Comunitaria*, *12*, 620-630.

Martínez de la Iglesia, J., Onís, M. C., Dueñas, H. R., Albert, C. C., Aguado, T. C., Colomer, A., Arias, C., Blanco, M. C. (2005). Abbreviating the brief. Approach to ultra-short versions of the Yesavage questionnaire for the diagnosis of depression. *Atención Primaria*, *35*(1), 14-21. https://doi.org/10.1157/13071040

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

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

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

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

Muñiz, J., Hernández, A., & Fernández-Hermida, J. R. (2020). Utilización de los test en España: El punto de vista de los psicólogos [Test use in Spain: the psychologists' viewpoint]. *Papeles del Psicólogo/Psychologist Papers*, *41*(1), 1-15. https://doi.org/10.23923/pap.psicol2020.2921

Ortega Orcos, R., Fort, M. S., Khajoui, A. K., Aparicio, S. V., & Valle, R. D. D. del (2007). Validación de la versión española de 5 y 15 ítems de la Escala de Depresión Geriátrica en personas mayores en Atención Primaria [Validation of the 5- and 15-item Spanish version of the Geriatric Depression Scale in older people in Primary Care]. *Revista Clínica Española*, *207*(11), 559-562. https://doi.org/10.1016/S0014-2565(07)73477-X

Ortuño-Sierra, J., Fonseca-Pedrero, E., Inchausti, F., & Sastre i Riba, S. (2016). Evaluación de dificultades emocionales y comportamentales en población infanto-juvenil: El cuestionario de capacidades y dificultades (SDQ) [Assessing behavioural and emotional difficulties in the child-adolescent population: the strengths and difficulties questionnaire (SDQ)]. *Papeles del psicólogo/Psychologist Papers*, *37*(1), 14-26. https://www.papelesdelpsicologo.es/pdf/2658.pdf

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

Pérez-Sánchez, J., Delgado, A. R., & Prieto, G. (2022). Evaluación del Emotion Regulation Checklist para Niños y Adolescentes [Evaluation of the Emotion Regulation Checklist for Children and Adolescents]. *Psicologia: Teoria e Pesquisa*, *38*. https://doi.org/10.1590/0102.3772e38213.es

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

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

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

Reyes-Pérez, Á., López-Martínez, A. E., Esteve, R., & Ramírez-Maestre, C. (2023). Spanish validation of the COMM Scale to assess the misuse of prescription opioids in patients with chronic noncancer pain. *International Journal of Mental Health and Addiction*, *21*(5), 3458-3472. https://doi.org/10.1007/s11469-022-00803-3

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

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

Schittekatte, M., & Evans, N. (2023, September 27). Updating the EFPA BoA Test Review Model: a necessary titanic work with many angles and supported by even more shoulders. [Conference object] Symposium at European Congress of Psychology 2023, Brighton. https://doi.org/10.23668/psycharchives.13272

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

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

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

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

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

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

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