Papeles del Psicólogo, 2011. Vol. 32(2), pp. 113-128
http://www.cop.es/papeles

REVIEW OF TESTS PUBLISHED IN SPAIN

José Muñiz[1], José R. Fernández-Hermida[1], Eduardo Fonseca-Pedrero[2], Ángela Campillo-Álvarez[1] and Elsa Peña-Suárez[1]
[1]*University of Oviedo.* [2]*University of La Rioja*

The proper use of psychological tests requires that the measurement instruments have adequate psychometric properties, such as reliability and validity, and that the professionals who use the instruments have the necessary expertise. In this paper we present the first review of tests published in Spain, carried out with an Assessment Model developed by the European Test Committee, and adapted to the Spanish context. The model permits both qualitative and quantitative assessment of the test. Ten tests were reviewed, selected from among those most widely used by Spanish professionals. Each test was sent to two peer reviewers for its assessment, and based on this assessment a final report was drawn up. In general, it can be said that the quality of the ten measurement instruments is good, the reports highlighting their strong and weak points. In light of the reviews some improvements are suggested for future editions of the tests, emphasizing the need to include in the Manuals as much evidence as possible on the validity of the tests. Finally, we discuss the details of the review process and analyze possible future directions for test assessment in Spain.
Key words: *Tests, Test use, Test assessment, Psychometrics.*

La utilización correcta de los tests psicológicos requiere por un lado que los instrumentos de medida tengan las propiedades psicométricas adecuadas, tales como fiabilidad y validez, y por otro, que los profesionales que los utilizan tengan la preparación técnica necesaria para usarlos. En el presente trabajo se presentan las primeras evaluaciones de tests editados en España, llevadas a cabo con un Modelo de Evaluación desarrollado por la Comisión Europea de Tests y adaptado al contexto español. El modelo permite llevar a cabo una evaluación tanto cualitativa como cuantitativa de las pruebas. Se evaluaron diez tests elegidos de entre los más utilizados por los profesionales españoles, cada uno de ellos se envió a dos revisores expertos para su evaluación, y a partir de dichos informes se elaboró el informe final. En líneas generales puede afirmarse que la calidad de los diez instrumentos de medida evaluados es buena, poniéndose de manifiesto sus puntos fuertes y débiles. A la vista de las revisiones se recomienda una mejora de las pruebas y sus Manuales para futuras ediciones, haciendo hincapié en la necesidad de incluir el mayor número posible de evidencias de validez sobre las pruebas. Finalmente se comentan los detalles del proceso de revisión seguido, y se analizan las posibles líneas de futuro en la evaluación de los tests en España.
Palabras clave: *Tests; uso de los tests; evaluación de tests; psicometría*

The correct use of measurement instruments in any professional field – and psychology is no exception – requires, on the one hand, that the instruments have adequate measurement properties, such as reliability and validity, and on the other, that the professionals using them have the necessary technical expertise to do so. An instrument with good psychometric properties can be rendered disastrous if the person using it is unqualified. One might almost say, paraphrasing the classic expression, that it seems the best tests are destined to fall into the hands of the worst users. The professional associations and various national and international organizations have been striving for several years to try and improve these two aspects – the quality of tests themselves and the competence of the professionals who use them. An account of such activities and projects can be consulted in Muñiz and Bartram (2007) or Muñiz and Fernández-Hermida (2010). Naturally, guaranteeing the correct use of tests is a necessary, but not sufficient, condition for the success of the entire psychological assessment process (Fernández-Ballesteros, De Bruyn, Godoy, Hornke, Ter Laak, & Vizcarro, 2001).

One of the most common demands from professional psychologists on expressing their opinions about test use concerns the need for the availability of technical information on tests to help them make the appropriate decisions (Evers, Muñiz, Bartram et al., in press; Muñiz & Fernández-Hermida, 2000, 2010; Muñiz et al., 1999, 2001). In response to this demand from European psychologists, the Standing Committee on Tests and Testing of the European Federation of Psychologists'

Correspondence: José Muñiz. *Facultad de Psicología. Universidad de Oviedo. Plaza Feijoo, s/n. 33003 Oviedo. España.*
E-mail: jmuniz@uniovi.es

Associations (EFPA-SCTT) developed a test assessment model which it made available to professionals in the European countries. This model can be consulted at (http://www.efpa.eu/professional-development/tests-and-testing). In Spain, the model was adapted by Prieto and Muñiz (2000) and published in this journal (see Appendix 1). The basic characteristic of this European model with respect to previous models, such as those developed in England (Bartram, 1996, 1998) or in Holland (Evers, 2001a, 2001b), is that it permits an exhaustive assessment of the different psychometric properties of tests, as well as providing both a quantitative and a qualitative rating of them. This model has been used in several European countries for test assessment, notably England and Holland; in the latter country, indeed, all published tests have been assessment with this or with previous models (Evers et al., 2010).

TEST ASSESSMENT IN SPAIN

As mentioned above, in Spain the European test assessment model was adapted for the Spanish context (Prieto & Muñiz, 2000), but it had not been used systematically up to now. In 2010 the Tests Commission of the Spanish Psychological Association (*Colegio Oficial de Psicólogos*; COP) made a unanimous decision to initiate the test assessment process. To this end, 10 tests were selected, taking into account both the extent to which they were used by Spanish psychologists (Muñiz & Fernández-Hermida, 2000, 2010) and publishers' interest in subjecting their tests to this first assessment process. In accordance with these two criteria, the 10 tests listed in Table 1 were assessed.

TABLE 1
LIST OF TESTS REVIEWED

Tests reviewed	
WAIS-III	Wechsler Adult Intelligence Scale - III
WISC-IV	Wechsler Intelligence Scale for Children - IV
MCMI-III	Millon Clinical Multiaxial Inventory-III
MMPI-2-RF	Minnesota Multiphasic Personality Inventory-2 Restructured Form
16PF-5	Sixteen Personality Factor questionnaire, fifth edition
PROLEC-R	*Batería de Evaluación de procesos Lectores, revisada* (Battery for the asssessment of reading procesess, revised)
EFAI	*Evaluación Factorial de la Aptitudes Intelectuales* (Factorial assessment of intellectual abilities)
NEO PI-R	Revised NEO Personality Inventory
EVALUA	*Batería Psicopedagógica* (EVALUA psychopedagogical battery)
IGF	*Batería de Inteligencia General y Factorial* (General and factorial intelligence battery)

ASSESSMENT PROCESS

Once the 10 tests had been selected for review, a pairwise assessment process was followed, similar to that used for reviewing scientific research articles and projects. The COP Tests Commission selected a set of reviewers, and each test was sent to two of them. Ideally, this pair of reviewers consisted of one whose expertise was of a more technical-psychometric nature, and another with more of a background in the substantive aspects of the variables measured by the test. This balance was achieved, if not for all the tests, in the majority of cases. Table 2 provides a list of the 20 reviewers who assessed the 10 tests.

The publishers provided two sets of each test free of charge, which were sent to the corresponding reviewers. Once the assessment was over, the tests were donated to the reviewers, who were also paid a symbolic 50 euros for their work. The response of the reviewers to whom the tests were sent can be considered exceptional, the number of rejections of the invitation being minimal, and always for reasons of *force majeure*. From here, and on behalf of the COP Tests Commission, we should like to express our sincere thanks for their contributions; none of this could have been done without their help. Once all the assessments had been returned, the Psychometrics group

TABLE 2
REVIEWERS WHO CARRIED OUT THE TEST ASSESSMENTS

Reviewer	Affiliation
María Victoria del Barrio Gándara	UNED (University of Distance Education)
Elisardo Becoña	University of Santiago de Compostela
María José Blanca Mena	University of Málaga
Isabel Calonge	Complutense University of Madrid
Antonio Cano Vindel	Complutense University of Madrid
Eduardo Fonseca Pedrero	University of La Rioja
María Forns	University of Barcelona
Jesús Enrique de la Fuente Arias	University of Almería
Olaya García	University of Barcelona
Juana Gómez Benito	University of Barcelona
Héctor González Ordi	Universidad Complutense
María Dolores Hidalgo Montesinos	University of Murcia
Serafín Lemos Giráldez	University of Oviedo
José Antonio López Pina	University of Murcia
Carmen Moreno	UNED
José Luis Miralles	University of Valencia
María José Navas	UNED
José Carlos Núñez	University of Oviedo
Vicente Ponsoda	Autónoma University of Madrid
Celestino Rodríguez	University of Oviedo

at the University of Oviedo, under the coordination of José Muñiz, drew up a joint report taking into account the assessments of both reviewers of the pair. As is the case with the assessment of scientific articles or research projects, this report is more than the mere sum of the reviewers' reports; rather, their appraisals are carefully considered to produce a report that reflects their opinions as faithfully as possible. In no case was it necessary to send the test to a third reviewer, since although in some instances there were certain discrepancies and differences of emphasis, they were able to be resolved satisfactorily. When the final report was finished it was sent to the publishers so that they and the authors could have the opportunity to express their point of view. The publishers' and authors' responses were highly professional, making it possible to clarify some aspects that were not sufficiently clear in the reviewers' reports. In our view, this step of providing the authors and publishers with the opportunity to give their opinion is fundamental, and for two reasons: on the one hand they are able to see at first hand the strengths and weaknesses of their tests, which helps them become aware of the need, where applicable, to modify some aspects of the instrument in subsequent versions; and on the other, it means that attention can be drawn to some details or features that the reviewers may have overlooked. Naturally, taking into account the opinion of the publishers and authors does not mean that the report will be modified simply because there may be some discrepancies, but it does permit the correction and clarification of some of the reviewers' appraisals. It should be stressed that the test assessment process does not constitute a "settling of scores" with the publishers and authors: the basic aim is to highlight the strong and weak points of the tests with a view to improving subsequent editions. Tests are living instruments, not written in stone, and the idea is that later versions can contribute evidence of validity that makes them more consistent and rigorous.

Table 3 offers a summary of the assessments of the 8 tests for which the reports were published. As it can be seen, in general the tests have a more than reasonable level of quality, all presenting some strong and some weaker points. The full reviews can be consulted on the COP website: www.cop.es, Tests Commission Section.

SOME LESSONS LEARNED

Without going into too much detail on each test, some general recommendations can be made. The first of these concerns the need to improve the Manuals, since they constitute the cornerstone supporting the contribution of evidence of the tests' validity. The manuals necessary today are far removed from the type of simple sheet provided in the past, which featured the norms and little else. It could indeed be said that a test is only as good as its manual, which should reflect all the evidence and data

TABLE 3
SUMMARY OF THE RATINGS AWARDED TO THE REVIEWED TESTS

	Characteristics							
	WISC-IV	**EVALUA**	**MMPI-2-RF**	**16PF**	**PROLEC-R**	**EFAI**	**NEO-PI-R**	**IGF**
Quality of materials and documentation	5	3.5	5	4.5	5	4.5	4	3
Theoretical foundations	4.5	2.5	5	4.5	5	4	3.5	3.5
Spanish adaptation	4.5	-	4	3	-	-	5	-
Item analysis	5	3.5	-	3	-	4.5	3.5	2
Content validity	5	3	4.5	4	5	4	4	3.5
Construct validity	3	4	4.5	4	4	4	3	3
Analysis of bias	-	-	-	-	-	-	-	-
Predictive validity	4	-	4	-	3	4	-	3
Reliability: equivalence	-	-	-	-	-	-	-	4
Reliability: internal consistency	4	3	4.5	3.5	3	5	4.5	3
Reliability: stability	3.5	-	4		-	-	-	-
Norms	4	4.5	4	4	3	5	4	3.5

Note. During the assessment period the commercial situation of the WAIS-III and MCMI-III changed, and given that the publishers have no guarantee of the rights over the tests, they requested that the reports were not published.
Scores in this table are on a scale of 1 to 5: 1 = inadequate; 2 = adequate but with shortcomings; 2.5 and over = adequate; 3.5 and over = good; 4.5 and over = excellent. A dash (-) signifies that no informaiton was provided.

relative to the instrument, as well as providing an updated list of references for it. For example, an aspect lacking in the majority of the manuals analyzed is an explicit approach to content validity, that is, explaining how it was guaranteed that the test contains an adequate representation of the items for assessing the construct in question. This does not mean that the tests analyzed lack content validity, but it does imply that insufficient effort has been made to set out clearly and directly the strategy followed for guaranteeing content validity, which is sometimes taken for granted. Likewise, in none of the tests reviewed were systematic analyses carried out on Differential Item Functioning, or bias, and this is a clearly improvable aspect; it is to be hoped that in future editions this type of analysis will be incorporated. It must be ensured that the test items function in a similar way for the different groups assessed, such as men and women, different age groups, or individuals from different backgrounds. In sum, the manuals, and consequently the tests, must begin to incorporate the features resulting from advances in psychometrics (Abad, Olea, Ponsoda, & García, 2011; AERA, APA, NCME, 1999; Bartram & Hambleton, 2006; Bennett, 2006; Downing & Haladyna, 2006; Drasgow, Luecht, & Bennett, 2006; Wilson, 2005).

Another aspect in which there is room for improvement is that, in the case of tests adapted from other countries, mostly the USA, they do not include exhaustive accounts of the evidence of validity already obtained in the country of origin. Not that this would exempt them from obtaining such evidence in Spanish population, but it would be a contribution to the accumulation of data on the instrument. Likewise, in some manuals there is no detailed account of the process of translation-adaptation of the test, nor information about the equivalence – if there is any – between the original forms and the adapted ones (Hambleton, Merenda, & Spielberger, 2005; Muñiz & Hambleton, 1996; van de Vijver & Hambleton, 1996).

In sum, a measurement instrument permits professionals to make inferences from scores obtained by individuals taking the test, so that manuals must provide in a detailed and rigorous way the evidence guaranteeing that those inferences can be made reliably and validly. It is true that each test has its own characteristics and peculiarities, but in all cases professionals should be informed which inferences are supported by empirical evidence and which are not.

LOOKING TO FUTURE ASSESSMENTS

What we present here are the beginnings of systematic test reviewing in Spain. It can be said that the experience is a clearly positive one and that the process should continue steadily; the destination is that one day all the tests published in this country will be assessed, as currently occurs in Holland. We continue by mentioning some of the lessons learned in this first attempt, in the hope that they prove useful for improving assessment practice in the future. A matter that has given rise to some discussion is the question of what is the best way of drawing up the final report on the test based on the reviewers' assessments. There is no single and unequivocal solution; the way it was approached here works well, and we are reasonably satisfied, but there are other possible options. Responsibility for the report could be given to a guide who is expert in the test, and who would combine the reports of the reviewers and resolve any discrepancy, given his or her knowledge of the instrument. This is the approach followed in England. For their part, the Dutch, with almost 30 years' experience (Evers et al., 2010), get the reviewers to interact until they arrive at an agreement about the test. We will need to decide which of the models to follow in the future; indeed, the best option may be to use a combination of them.

Another aspect to be elucidated, looking to the future, is which tests to choose for continuing the assessment. In the case of these first 10 they were selected on the basis of agreement by the COP Tests Commission, but perhaps in future publishers could be invited to submit for review those instruments they consider appropriate, though it is important for the Tests Commission to ensure that the instruments selected are of sufficient relevance for professionals.

As for the European Test Assessment Model, on which the model used here is based (Appendix 1), it is currently under review by a European committee. Once the new version is available, we shall incorporate the corresponding modifications in our model. The most urgent issues for consideration, and on which the committee is working, are remote assessment via Internet, Automated Reports, the technology of Item Response Theory, Computerized Tests, and everything related to Criterion-Referenced Tests. It is important to include these aspects in the model, especially as the review process begins to cover tests developed with new psychometric technologies. While for reviewing the most classical tests the failure to consider these techniques is not a problem –

since the instruments themselves do not include them –, as the assessment net widens it will surely be necessary to review tests that do include them, and therefore to have an appropriate model.

As regards the application of the model in practice, no serious problems were detected by the reviewers, even though some aspects could be considered for improvement. Thus, for example, some reviewers do not carry out the general ratings of the tests in quantitative fashion as required in the table designed for this purpose. Apparently, the instructions are not sufficiently clear, so that it would be useful to stress that the ratings should be made based on the calculation of the arithmetic mean of the scores assigned in the different sections specified in the general ratings table. Other areas where there is room for improvement concern confusions over concepts, for instance in section 1.15. There is also some confusion over determining whether or not there are different forms of the test (section 1.18). Some reviewers include in this section the possibility of obtaining computerized reports, even though the Assessment Model refers to whether there exist parallel forms, brief versions or computerized versions.

Likewise, it would be appropriate to reformulate the section on scoring (section 1.19), or to explain what each of its categories refers to; at times, *Optical reader* and *Automated via computer* are confused. Reviewers tend to understand that the optical reader constitutes a scoring method automated via the computer. As for content that might be included in future versions of the Test Review Questionnaire, an important aspect is the general description of the test. It might be useful to request details of all the revisions carried out since the first publication of the test assessed (section 1.9) and to stress the need for a description of the scales making up the reviewed test, given that some reviewers merely list them. Even when there are several subscales or sections it would be advantageous to specify the number of items in each one of them (section 1.14). It would also be appropriate to reformulate in quantitative fashion the item rating the basic bibliography provided in the manual. The question could be framed in these terms: *Please rate the basic bibliography about the test provided in the documentation as: inadequate (1); adequate but with shortcomings (2); sufficient (3); good (4); excellent (5).*

Other suggestions for improvements would involve enlarging the section on the adaptation of the test. Given the relevance and importance of the process of translation and adaptation of tests from abroad, which in many cases is inadequate, it would be advantageous to obtain more information about this process. For example, items could be included about adaptation methods such as back-translation or double translation, what type of professionals have carried out these tasks, whether they followed the appropriate international guidelines, and so on. These and other more general questions already mentioned should be taken into account on preparing a new version of the Test Review Questionnaire.

ACKNOWLEDGEMENTS

We should like to express our sincere thanks to the members of the Tests Commission of the Spanish Psychological Association, without whose help this work would not have been possible. Many thanks to Eduardo Montes Velasco, Rocío Fernández Ballesteros, Miguel Martínez García, Jaime Pereña Brand and Javier Rubio Ramiro.

REFERENCES

Abad, F. J., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en ciencias sociales y de la salud* [*Measurement in the social and health sciences*]. Madrid: Síntesis.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Bartram, D. (1996). Test qualifications and test use in the UK: The competence approach. *European Journal of Psychological Assessment, 12*, 62-71.

Bartram, D., & Hambleton, R. K. (Eds.) (2006). *Computer-based testing and the Internet.* Chichester, UK: Wiley and Sons.

Bennett, R. E. (2006). Inexorable and inevitable: The continuing story of technology and assessment. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the Internet* (pp. 201-217). Chichester, UK: Wiley and Sons.

Downing, S. M., & Haladyna, T. M. (Eds.) (2006). *Handbook of test development.* Hillsdale, NJ: Erlbaum.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (pp. 471-515).Westport, CT: ACE/Praeger.

Evers, A. (2001a). Improving test quality in the Netherlands: Results of 18 years of test ratings. *International Journal of Testing, 1*, 137-153.

Evers. A. (2001b). The revised Dutch rating system for test quality. *International Journal of Testing, 1*, 155-182.

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., Frans, O., Gintiliené, G., Hagemeister, C., Halama, P., Iliescu, D., Jaworowska, A., Jiménez, P., Manthouli, M., Matesic, K., Schittekatte, M., Sümer, C., & Urbánek, T. (in press). Testing practices in the 21st century: Developments and European psychologists' opinions. *European Psychologist*.

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.

Eyde, L. D., Robertson, G. J., Krug, S. E., Moreland, K. L., Robertson, A. G., & Shewan, C. M., et al. (1993). *Responsible test use. Case studies for assessing human behavior.* Washington, DC: American Psychological Association.

Fernández-Ballesteros, R., De Bruyn, E., Godoy, A., Hornke, L., Ter Laak, J., Vizcarro, C., et al. (2001). Guidelines for the assessment process (GAP): A proposal for discussion. *European Journal of Psychological Assessment, 17*, 187-200.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment.* London: Erlbaum.

Muñiz, J., & Bartram, D. (2007). Improving international tests and testing. *European Psychologist, 12*, 206-219.

Muñiz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., Fernández-Hermida, J.R., & Zaal, J. (2001). Testing practices in European countries. *European Journal of Psychological Assessment, 17*, 201-211.

Muñiz, J., & Fernández-Hermida, J.R. (2000). La utilización de los tests en España [Test use in Spain]. *Papeles del Psicólogo, 76*, 41-49.

Muñiz, J., & Fernández-Hermida, J. R. (2010). La opinión de los psicólogos españoles sobre el uso de los tests [The opinion of Spanish psychologists on the use of tests]. *Papeles del Psicólogo, 31(1),* 108-121.

Muñiz, J., & Hambleton, R. K. (1996). Directrices para la traducción y adaptación de los tests [Guidelines for the translation and adaptation of tests]. *Papeles del Psicólogo*, 66, 63-70.

Muñiz, J., Prieto, G., Almeida, L., & Bartram, D. (1999). Test use in Spain, Portugal and Latin American countries. *European Journal of Psychological Assessment, 15*, 151-157.

Prieto, G., & Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España [A model for assessing the quality of tests used in Spain]. *Papeles del Psicólogo, 77*, 65-71.

van de Vijver. F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89-99.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Erlbaum.

APPENDIX 1
QUESTIONNAIRE USED FOR TEST REVIEW

1. General description of the test[1]

1.1. Name of test:
1.2. Name of test in its original version (if the Spanish version is an adaptation):
1.3. Author(s) of original test:
1.4. Author(s) of Spanish adaptation:
1.5. Publisher of original version:
1.6. Publisher of Spanish adaptation:
1.7. Publication date of original test:
1.8. Publication date of Spanish adaptation:
1.9. Latest test revision of Spanish adaptation:
1.10. Indicate the general area of the variable or variables to be measured by the test[2]

( ) Intelligence
( ) Abilities/Aptitudes
( ) Skills and academic performance
( ) Psychomotor skills
( ) Neuropsychology
( ) Personality
( ) Motivation
( ) Attitudes
( ) Interests
( ) Development scales
( ) Curricular competence
( ) Clinical scales
( ) Learning potential
( ) Others (Specify:.....)

1.11. Brief description of the variable or variables to be measured by the test:
(A non-judgemental description (around 200-600 words) of the test. This description should give the reader a clear idea of the test, what it sets out to measure and its scales)

1.12. Field of application[3]

( ) Clinical Psychology
( ) Educational Psychology
( ) Neuropsychology
( ) Forensic Psychology
( ) Work and Organizational Psychology
( ) Sports Psychology
( ) Social services
( ) Traffic Psychology
( ) Others (Specify:.....)

[1] If the test is made up of subtests that are heterogeneous in format and characteristics, please fill out a different questionnaire for each subtest.
[2] More than one option can be marked.
[3] More than one option can be marked.

1.13. Item Format[4]:
( ) Free response
( ) Dichotomous response (yes/no, true/false, etc.)
( ) Multiple-choice
( ) Likert-type
( ) Bipolar adjectives
( ) Other (Specify :.....)

1.14. Number of items[5]:

1.15. Administration mode[6]:
( ) Oral administration
( ) Pencil and paper
( ) Manipulative
( ) Computerized
( ) Other (Specify :.....)

1.16. Qualifications needed for using the test according to the documentation presented:
( ) None
( ) Training and specific accreditation*
( ) Level A[7]
( ) Level B
( ) Level C
( ) Other (Specify :.....)

*Indicate name of institution providing the accreditation:

1.17. Description of populations to which the test is applicable (specify age range, educational level, etc., and whether the test is applicable in certain specific populations: ethnic minorities, special needs, clinical groups, etc.):

1.18. Specify whether there are different forms of the test and their characteristics (parallel forms, brief versions, computerized or printed versions, etc.) In the case that there are computerized versions, specify the minimum hardware and software requirements.

1.19. Scoring procedure:
( ) Manual using template
( ) Optical reader
( ) Computerized
( ) Scored exclusively by the test supplier
( ) Scored by experts
( ) Self-correctable
( ) Other (Specify:.....).

[4] More than one option can be marked.
[5] If the test has several scales, specify number of items on each one.
[6] More than one option can be marked.
[7] Some countries have adopted systems for classifying tests in different categories, according to the qualification required by users. These classification systems provide test publishers with a means of deciding to whom they might sell the test. A very widely used system is that which divides tests in three categories: Level A (tests of performance and knowledge), Level B (collective tests of abilities and intelligence) and Level C (individually-applied tests of intelligence or personality and other complex instruments).

1.20. Scores: (Describe the procedure for obtaining raw scores).

1.21. Transformation of scores:
( ) Not applicable to this instrument
( ) Normalized
( ) Non-normalized

1.22. Scales used:
( ) Centile
( ) Standard scores
( ) Deviation ratios
( ) Enneatypes
( ) Decatypes
( ) T (Mean 50 and standard deviation 10)
( ) S (Mean 50 and standard deviation 20)
( ) Other (Specify:.....)

1.23. Possibility of obtaining automatized reports
( ) No
( ) Yes*

*In case of an affirmative answer, give a brief non-judgemental description of the Automatized Report, indicating the basic characteristics, such as type of report, structure, clarity, style, tone, etc.

1.24. Does the publisher offers a service for scoring and/or generating reports?:
( ) No
( ) Yes

1.25. Estimated time for application of the test (instructions, examples and responses to items).
Individual application:.....

Collective application:.....

1.26. Documentation provided by the publisher:
( ) Manual
( ) Books or complementary articles
( ) Discs/CD
( ) Other (Specify :.....)

1.27. Price of a complete set of the test (documentation, test, scoring templates; in the case of computerized tests the price of the hardware is not included):

1.28. Price and number of copies of the pack of booklets (pencil and paper tests):

1.29. Price and number of copies of the pack of response sheets (pencil and paper tests):

1.30. Price of scoring and/or generated reports by publisher:

1.31. Basic bibliography about the test provided in the documentation:

2. Rating of test characteristics

2.1. Quality of test materials (objects, printed material or software):

* ( ) Inadequate
** ( ) Adequate but with some shortcomings
*** ( ) Adequate
**** ( ) Good
***** ( ) Excellent (High-quality printing and presentation, very attractive and efficient software, etc.)

2.2. Quality of the documentation provided:

* ( ) Inadequate
** ( ) Adequate but with some shortcomings
*** ( ) Adequate
**** ( ) Good
***** ( ) Excellent (Very clear and complete description of the technical characteristics, founded on abundant data and references)

2.3. Theoretical foundations:

( ) No information provided in the documentation
* ( ) Inadequate
** ( ) Adequate but with some shortcomings
*** ( ) Adequate
**** ( ) Good
***** ( ) Excellent (Very clear description of the construct to be measured and of the measurement process)

2.4. Adaptation of the test (if the test has been translated and adapted for its application in Spain):

( ) Not applicable to this instrument
( ) No information provided in the documentation
* ( ) Inadequate
** ( ) Adequate but with some shortcomings
*** ( ) Adequate
**** ( ) Good
***** ( ) Excellent (Precise description of the translation procedure, of the adaptation of the items to the Spanish culture, of studies of equivalence with the original version, use of International Test Commission guidelines, etc.).

2.5. Quality of the instructions:

* ( ) Inadequate
** ( ) Adequate but with some shortcomings
*** ( ) Adequate
**** ( ) Good
***** ( ) Excellent (Clear and precise. Highly adequate for the populations to which the test is addressed).

2.6. Ease of understanding of the task:

* ( ) Inadequate
** ( ) Adequate but with some shortcomings
*** ( ) Sufficient
**** ( ) Good
***** ( ) Excellent (Individuals from the populations to which the test is addressed can easily understand the task requirements).

2.7. Ease of recording responses:
* ( ) Inadequate
** ( ) Adequate but with some shortcomings
*** ( ) Adequate
**** ( ) Good
***** ( ) Excellent (The procedure for giving responses is very simple, which helps in the avoidance of errors in recording them).

2.8. Item quality (formal aspects):
* ( ) Inadequate
** ( ) Adequate but with some shortcomings
*** ( ) Adequate
**** ( ) Good
***** ( ) Excellent (Wording and design are highly appropriate)

2.9. Item analysis

2.9.1 Data on item analysis:
( ) Not applicable to this instrument
( ) No information provided in the documentation
* ( ) Inadequate
** ( ) Adequate but with some shortcomings
*** ( ) Adequate
**** ( ) Good
***** ( ) Excellent (Detailed information on diverse studies about the psychometric characteristics of the items: difficulty or variability, discrimination, validity, distractors, etc.)

2.10. Validity

2.10.1. Content validity[8]:

2.10.1.1. Quality of the representation of the content or domain:
***** ( ) Inadequate
***** ( ) Adequate but with some shortcomings
***** ( ) Adequate
***** ( ) Good
***** ( ) Excellent (The documentation includes a precise definition of the content. The items adequately sample all the facets of the content)

2.10.1.2. Consultations with experts[9]:
( ) No information provided in the documentation
* ( ) No experts were consulted about the content representation
** ( ) Experts were consulted only informally or in small number
*** ( ) A small number of experts were consulted using a systematic procedure (N<10)
**** ( ) A moderate number of experts were consulted using a systematic procedure (10≤N≤30)
***** ( ) A large number of experts were consulted using a systematic procedure (N>30)

[8]This aspect is essential in criterion-referenced tests, and particularly in academic performance tests. Make your judgement about the quality of the representation of the content or domain. If the documentation provided includes the experts' assessments, take them into consideration.

[9] The figures on sample sizes and statistics that appear below are for guidance only.

2.10.2. Construct validity:

2.10.2.1. Designs employed10:
( ) No information provided in the documentation
( ) Correlations with other tests
( ) Differences between groups
( ) Multitrait-multimethod matrix
( ) Exploratory factor analysis
( ) Confirmatory factor analysis
( ) Experimental designs
( ) Others (Specify:.....).

2.10.2.2. Sample sizes in the construct validation:
( ) No information provided in the documentation
* ( ) One study with a small sample (N<200)
** ( ) One study with a moderate sample (200≤N≤500)
*** ( ) One study with a large sample (N>500)
**** ( ) Several studies with moderate samples
***** ( ) Several studies with large samples

2.10.2.3. Sample-selection procedure*:
( ) No information provided in the documentation
( ) Incidental
( ) Random
*Briefly describe the selection procedure.

2.10.2.4. Median of the correlations of the test with similar tests:
( ) No information provided in the documentation
* ( ) Inadequate (r<0.25)
** ( ) Adequate but with some shortcomings (0.25≤r<0.40)
*** ( ) Adequate (0.40≤r<0.50)
**** ( ) Good (0.50≤r<0.60)
***** ( ) Excellent (r≥0.60)

2.10.2.5. Quality of the tests employed as criterion or marker:
( ) No information provided in the documentation
* ( ) Inadequate
** ( ) Adequate but with some shortcomings
*** ( ) Adequate
**** ( ) Good
***** ( ) Excellent

2.10.2.6. Data on item bias
( ) Not applicable to this instrument
( ) No information provided in the documentation
* ( ) Inadequate
** ( ) Adequate but with some shortcomings
*** ( ) Adequate

[10] Puede marcar más de una opción.

**** ( ) Good
***** ( ) Excellent (Detailed information on diverse studies about item bias related to sex, mother tongue, etc. Use of appropriate methodology)

10 More than one option can be marked.

2.10.3. Predictive validity

2.10.3.1. Describe the criteria employed and the population characteristics:

2.10.3.1. Design of criterion selection11:
( ) Concurrent
( ) Predictive
( ) Retrospective

2.10.3.2. Sample sizes in the predictive validation:
( ) No information provided in the documentation
* ( ) One study with a small sample (N<100)
** ( ) One study with a moderate sample (100≤N<200)
*** ( ) One study with a large and representative sample (N≥200)
**** ( ) Several studies with moderate and representative samples
***** ( ) Several studies with large and representative samples

2.10.3.3. Sample-selection procedure*:
( ) No information provided in the documentation
( ) Incidental
( ) Random

*Briefly describe the selection procedure.

2.10.3.4. Median of correlations of the test with the criteria:
( ) No information provided in the documentation
* ( ) Inadequate (r<0.20)
** ( ) Sufficient (0.20≤r<0.35)
*** ( ) Good (0.35≤r<0.45)
**** ( ) Very good (0.45≤r<0.55)
***** ( ) Excellent (r≥0.55)

2.10.4. Comments on the validity in general:

2.11. Reliability

2.11.1. Data provided on reliability:
( ) A single reliability coefficient
( ) A single standard error of measurement
( ) Reliability coefficients for different groups of respondents
( ) Standard error of measurement for different groups of respondents

[11] More than one option can be marked.

2.11.2. Equivalence (Parallel forms):
2.11.2.1. Sample sizes in the equivalence studies:
( ) No information provided in the documentation
* ( ) One study with a small sample (N<200)
** ( ) One study with a moderate sample (200≤N<500)
*** ( ) One study with a large sample (N>500)
**** ( ) Several studies with moderate samples
***** ( ) Several studies with large samples

2.11.2.2. Median of the equivalence coefficients:
( ) No information provided in the documentation
* ( ) Inadequate (r<0.50)
** ( ) Adequate but with some shortcomings (0.50≤r<0.60)
*** ( ) Adequate (0.60≤r<0.70)
**** ( ) Good (0.70≤r<0.80)
***** ( ) Excellent (r≥0.80)

2.11.3. Internal consistency

2.11.3.1. Sample sizes in the consistency studies:
( ) No information provided in the documentation
* ( ) One study with a small sample (N<200)
** ( ) One study with a moderate sample (200≤N<500)
*** ( ) One study with a large sample (N≥500)
**** ( ) Several studies with moderate samples
***** ( ) Several studies with large samples

2.11.3.2. Median of the consistency coefficients:
( ) No information provided in the documentation
* ( ) Inadequate (r<0.60)
** ( ) Adequate but with some shortcomings (0.60≤r<0.70)
*** ( ) Adequate (0.70≤r<0.80)
**** ( ) Good (0.80≤r<0.85)
***** ( ) Excellent (r≥0.85)

2.11.4. Stability (Test-Retest)

2.11.4.1. Sample sizes in the stability studies12:
( ) No information provided in the documentation
* ( ) One study with a small sample (N<100)
** ( ) One study with a moderate sample (100≤N<200)
*** ( ) One study with a large sample (N≥200)
**** ( ) Several studies with moderate samples
***** ( ) Several studies with large samples

[12] Number of respondents with both scores (before-after).

2.11.4.2. Median of the stability coefficients:
( ) No information provided in the documentation
* ( ) Inadequate (r<0.55)
** ( ) Adequate but with some shortcomings (0.55≤r<0.65)
*** ( ) Adequate (0.65≤r<0.75)
**** ( ) Good (0.75≤r<0.80)
***** ( ) Excellent (r≥0.80)

2.11. 5 Comments on the reliability in general:

2.12. Norms

2.12.1. Quality of the norms:
( ) No information provided in the documentation
* ( ) One norm which is not applicable to the target population
** ( ) One norm applicable to the target population with some precautions
*** ( ) One norm adequate for the target population
**** ( ) Several norms addressing diverse populational strata
***** ( ) A wide range of norms according to age, sex, educational level and other relevant characteristics

2.12.2. Sample sizes13:
( ) No information provided in the documentation
* ( ) Small (N<150)
** ( ) Sufficient (150≤N<300)
*** ( ) Moderate (300≤N<600)
**** ( ) Large (600≤N<1000)
***** ( ) Very large (N≥1000)

2.12.3. Sample-selection procedure*:
( ) No information provided in the documentation
( ) Incidental
( ) Random
*Briefly describe the selection procedure.

2.12.4. Comments on the norms

3. Global appraisal of the test

3.1. In no more than 1000 words, please give your appraisal of the test, highlighting its strong and weak points, as well as making recommendations about its use in various professional areas. Also, please indicate any characteristics of the test that could be improved, information lacking from the documentation, etc.

[13] If there are several norms, respond for the average size

By way of summary, please fill out Tables 1 and 2.

Table 1 includes some descriptive data of the test.

TABLE 1
DESCRIPTION OF THE TEST

Characteristic	Description
Name of test (section 1.1)	
Author (section 1.3)	
Author of the Spanish adaptation (section 1.4)	
Date of latest revision (section 1.9)	
Construct assessed (section 1.11)	
Areas of application (section 1.12)	
Administration mode (section 1.15)	

Table 2 summarizes the appraisal of the test's general characteristics. Take into consideration the average of the ratings awarded in the sections indicated in the second column of Table 2.

TABLE 2
APPRAISAL OF THE TEST

Characteristic	Sections	Rating
Materials and documentation	2.1 and 2.2	
Theoretical foundations	2.3	
Adaptation	2.4	
Item analysis	2.9	
Content validity	2.10.1	
Construct validity	2.10.2	
Bias analysis	2.10.2.6	
Predictive validity	2.10.3	
Reliability: equivalence	2.11.2	
Reliability: internal consistency	2.11.3	
Reliability: stability	2.11.4	
Norms	2.12	