Papeles del Psicólogo / Psychologist Papers, 2016. Vol. 37(3), pp. 192-197
http://www.papelesdelpsicologo.es
http://www.psychologistpapers.com

ASSESSING THE QUALITY OF TESTS IN SPAIN: REVISION OF THE SPANISH TEST REVIEW MODEL

Ana Hernández[1], Vicente Ponsoda[2], José Muñiz[3], Gerardo Prieto[4] and Paula Elosua[5]
[1]*Universidad de Valencia.* [2]*Universidad Autónoma de Madrid.* [3]*Universidad de Oviedo.*
[4]*Universidad de Salamanca.* [5]*Universidad del País Vasco*

Para usar adecuadamente los tests, es necesario que los profesionales cuenten con información rigurosa de su calidad. Es por ello que, desde hace unos años, se viene aplicando el modelo español de evaluación de la calidad de los tests (Prieto y Muñiz, 2000). El objetivo de este trabajo es actualizar y revisar dicho modelo, con el fin de incorporar las recomendaciones hechas en sus aplicaciones, y para incorporar los avances psicométricos y tecnológicos que se han producido durante los últimos años. El modelo original fue revisado en varias fases, y la revisión originalmente propuesta fue revisada por un conjunto de expertos, lo que dio lugar a la versión final que se describe en este trabajo. Se espera que la aplicación del modelo revisado y la publicación de los resultados correspondientes, contribuya a seguir mejorando el uso de los tests y, con ello, la práctica profesional de la Psicología.
Palabras clave: *Uso de tests, Modelos de evaluación, Calidad de los tests.*

In order for practitioners to be able to use tests appropriately, they must have rigorous information on the quality of the tests. This is why the Spanish test review model (Prieto & Muñiz, 2000) has been applied for a number of years. The goal of this paper is to update and revise this model in order to incorporate the recommendations that have been provided since the original model was applied and to incorporate the latest psychometric and technological innovations. The original model was revised following a series of steps, and the revised proposal was reviewed by a number of experts. After incorporating their suggestions, we have arrived at the final version, which is described in this paper. With the application of the revised model, and the publication of the corresponding results, we hope to continue to improve the use of tests, and consequently, the professional practice of psychology.
Key words: *Test use, Review models, Test quality.*

It is a well-known fact that tests are a basic tool for the professional practice of psychology and they can be useful regardless of the area of professional expertise: social, educational, clinical, sports, legal, organizational, etc. In Spain, we find data that confirms that psychologists use the tests as a basic tool in their daily lives, when we use the survey designed by the EFPA (European Federation of Psychological Associations) to obtain the views of psychologists on the use of tests. When Spanish collegiate psychologists were asked about the frequency with which they used tests in their professional work, the means obtained in the different areas (clinical, organizational, educational psychology and others), were close to 4 on a 5 - point scale (Muñiz & Fernández-Hermida, 2010). Even higher scores were obtained in the items in which it is recognized that the tests are an excellent source of information when combined with other data, and which, when properly used, are a great help to the psychologist.

However, it is also known that for the tests to be a truly useful tool, they must have demonstrated quality and rigour. Moreover, psychologists, as the users of the tests must be competent and have proven information to help them choose tests with psychometric rigour for their purpose. In this sense, Spanish psychologists, in the aforementioned survey, reported needing more information (independent reviews, investigations, documentation, etc.) on the quality of the tests published in Spain.

It is in this context that the models for assessing the quality of tests and, in particular, the Spanish test review model (Prieto & Muñiz, 2000) arise. These models have in common that they define a set of criteria of theoretical, practical and / or psychometric quality, and they are evaluated following a standardized procedure in order to publicize the results of the evaluation subsequently. The ultimate goal is clear: to provide test users with accurate and accessible information on the quality of the tests available. Among the proposed models, and their application processes, we highlight the one followed by the American Buros Center for Testing and that proposed by the European Federation of Psychology Associations (EFPA), along with local models such as the Dutch, the British, and, of course, the Spanish model.

The Spanish model was driven by the Association of Psychologists (COP), and culminated in the publication of the CET (Test Assessment Questionnaire) in 2000 (Prieto & Muñiz, 2000). However, the CET was not implemented until years later, with the first test assessment process, which ended in 2011,

Correspondence: Ana Hernández. *IDOCAL y Departamento de Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Universitat de València. España.*
E-mail: Ana.Hernandez@uv.es

promoted by the COP and its test commission (see Muñiz, Fernández-Hermida, Fonseca-Pedrero, Campillo-Álvarez & Peña-Suarez, 2011). Since then, with some minor modifications, the model has been applied systematically, always by independent reviewers. The results were published on the COP's website. Summaries of the four evaluation processes carried out to date, and their main results have been collected in the work of Muñiz et al. (2011), Ponsoda and Hontangas (2013), Hernández, Tomás, Ferreres and Lloret (2014), and Elosua and Geisinger (2016).

As a result of these evaluations, and thanks to the experience gained in applying the model, a series of recommendations have been proposed and implemented to improve both the test assessment process and the model itself (see Elosua & Geisinger, 2016; Hernández et al, 2014; Muñiz et al, 2011; and Ponsoda & Hontangas, 2013). As regards the model, some of these recommendations have to do with the need to incorporate the psychometric and technological advances that have occurred in recent years. In fact, the European model of the EFPA has recently been revised and updated (Evers et al., 2013) in order to incorporate these advances. The goal is to bring them to the professional practice of psychologists and thereby help reduce the gap that frequently exists between research and professional practice (Elosua, 2012).

Given these considerations, the objective of this paper is to present the review and update of the CET. For this review, together with the recommendations made in the various evaluation processes where the CET was applied, the recently revised and updated EFPA model (Evers et al., 2013) was taken into account.

With the application of this revised model (CET-R), we hope to increase the clarity and wealth of information provided in the evaluation process. The subsequent publication of the results obtained with the revised model will help to disseminate more comprehensive and rigorous information on the quality of the tests and their weaknesses. With this we hope to continue contributing to the informative strategy initiated by the COP years ago, in order to improve the use of tests and, consequently, to improve the professional practice of psychology.

Firstly, we briefly present the main international test review models and, secondly, the original CET model is presented. Thirdly, the process followed in revising the model is described and the main innovations of the revised model (CET-R) are highlighted. Finally we close with some brief conclusions.

SOME INTERNATIONAL TEST REVIEW MODELS

The pioneers in systematically presenting information on the quality of tests were Americans through Buros, which is a testing institute associated with the University of Nebraska. In 1938 the first edition of the Buros Mental Measurements Yearbooks series was published with the results of the evaluations conducted. This series continues to be published regularly today (Buros, 1938; Carlson, Geisinger, & Jonson, 2014). Notably, for some years, and given the increase of Spanish speakers in the United States, Buros has special publications devoted to tests published in Spanish (see, for example, Carlson & González, 2015). The process and the characteristics of the Buros evaluation as well as the similarities and differences between the evaluation carried out by Buros and that which has been carried out so far in implementing the CET can be consulted in Elosua and Geisinger (2016).

Focusing on Europe, progress in this area has been mainly driven by the corresponding psychology associations, through their test commissions. The Dutch were the first to carry out systematic evaluations of the tests and publish the results of these evaluations. Specifically, the first publication was in 1969 (NIP, 1969). The model that was used then has been revised five times, the latest revision being carried out in 2009 (Evers, Braak, Frima, & Van Vliet-Mulder, 2009). A more detailed description of the history, process and results of the assessments carried out by the Dutch can be consulted in the article by Evers, Sijtsma, Lucassen and Meijer (2010). The Netherlands was followed by the British Psychological Society, albeit many years later. Although they began applying their own model in the 90s (see Bartram, 1996; Bartram, Lindley & Foster, 1990; Bartram, Lindley, & Marshall, 1992; Bartram, Anderson, Kellett, Lindley & Robertson, 1995; Bartram, Burke, Kandola, Lindley, Marshall, & Rasch, 1997), in recent years they have adopted the EFPA model, proposed in 2002 (see Bartram, 2002), which was based on local models proposed by the Dutch, British and Spanish. In addition to the British (e.g., Lindley, 2009), the EFPA model has been applied in recent years by Norwegians and Germans (see, for example, Nielsen, 2009 and Moosbrugger et al., 2009, respectively).

However, as mentioned above, the new developments in the field of psychological and educational evaluation have led to the EFPA model being thoroughly reviewed recently (see Evers et al., 2013). This revised version allows the comprehensive evaluation of tests. Similar to CET, firstly the test is described exhaustively, and secondly a quantitative assessment is performed of the psychometric characteristics of the test. In both parts, the descriptive and quantitative information is complemented with qualitative comments that enrich the evaluation. However, unlike the CET, the revised EFPA model includes sections to assess in detail aspects of the latest technological and psychometric advances: online test administration, the development of automated reports, and the application of Item Response Theory (IRT) among others.

The revised CET model, the CET-R, includes some of these new aspects, but not all of them. Only those considered most suitable for the Spanish context have been included, while attempts have also been made to maintain the parsimony of the CET model and facilitate the comparability between the new assessments and those carried out to date using the original model.

THE TEST EVALUATION QUESTIONNAIRE (CET): THE ORIGINAL MODEL AND THE APPLICATION PROCESS

The CET was designed primarily to evaluate tests constructed from classical test theory, and is structured in three sections. The first section, focused on the technical description of the test, contains 31 items referring to test name, author, construct

measured, application area, etc. The second section deals with the technical evaluation of the instrument characteristics. Items related to the quality of materials and the documentation, the instructions and items, the theoretical foundation, the adaptation/translation (if the test was originally constructed in another country), the analysis of the items, the study of the validity (differentiating the content, construct, and predictive validity and differential item functioning analysis (DIF)), the study of reliability (differentiating parallel forms, internal consistency and test retest), and the test norms. In total this section includes 32 closed items that are mostly answered by a response scale with five categories ranked according to the quality of the assessed characteristics. It also has several open items which request a reasoned justification of the responses to the closed items for each of the main characteristics evaluated (validity, reliability and test norms), as well as the description of the selection procedures of the samples used to evaluate the psychometric quality of the test, and the criteria used when evaluating the predictive validity. Finally, in the third and final section, an overall assessment of the test is requested and a summary of the first two assessment sections, which is presented in a data sheet.

The test evaluation process using the CET starts with the selection, by the COP's test commission, of both the tests to be evaluated and the coordinator who will manage the evaluation process. For each selected test, the coordinator chooses two reviewers, who work independently: one, an expert in psychometrics and the other an expert in the professional field of assessment on which the test is focused. The reviewers must not have a direct relationship with the authors of the tests, or express a conflict of interest that would call into question the objectivity of the assessment. The coordinator is responsible for integrating the evaluations of both reviewers into one final report. If there is no substantial agreement between the reviewers, a third one could be asked. The report generated is sent to the author and/or publisher of the test so they can make any observations and clarifications and provide additional information. Finally, after the appropriate modifications, the report is made public through the website of the COP.

In the first application of the model, Muñiz et al., (2011) highlighted the need to improve the instructions for completing the model, since not all of the evaluators seemed to follow the same criteria in responding to some of the items and some of them were not interpreted correctly. This need was confirmed by Ponsoda and Hontangas (2013) in the second evaluation. Therefore, and taking into account the suggestions made, from the third evaluation onwards additional instructions were provided to clarify in more detail what was expected from reviewers in responding to the questionnaire, in order to reduce ambiguities and standardize the process further (see Hernández et al., 2014).

Regarding the questionnaire itself, in the first two evaluation processes, recommendations on the inclusion of certain issues were noted. While many of the suggestions were not included in the following evaluations pending a more thorough review of the model which would take into account the revised and updated EFPA model –which we address in this work– clarifications were however included for some items and some new issues concerning construct validity, measurement precision using IRT and updating the test norms, among others (see details in Hernández et al., 2014). The additional instructions and minor changes to the CET were maintained in the fourth test evaluation (Elosua & Geisinger, 2016), which again highlighted the need for a more in depth review of the model.

THE NEW TEST EVALUATION QUESTIONNAIRE (CET-R): A DESCRIPTION OF THE MAIN CHANGES

Starting with the original CET with the small modifications applied by Hernandez et al. (2014), the first two authors worked on an initial proposal for CET-R which, on the one hand, would solve the problems of interpretation still observed in the assessments performed and, on the other, would incorporate some of the psychometric and technological advances made in recent years. We proceeded in four phases. First, we reviewed the suggestions made by the coordinators of the various test evaluation editions carried out by introducing the corresponding amendments and instructions. Second, we reviewed the updated model of the EFPA (Evers et al., 2013), adding the issues we considered most appropriate for the Spanish context, plus some others that we considered particularly relevant. Thus, in the initial proposal evaluation, assessment criteria were included on certain validation strategies, other ways of evaluating the reliability, and the interpretation of criterion-referenced test scores. However, we have left out the comprehensive assessment of issues such as the computerized administration of tests, remote evaluation via the Internet, or the quality of automated reports, although on the latter question an open item has been added to assess the quality of the report, in addition to maintaining the item that already existed describing the type of report. We have also left out a comprehensive assessment of the implementation of IRT (there are only two evaluative items related to the accuracy and adequacy of the sample size when IRT is applied), and continuous norming (although there is one question on this too). These aspects were excluded, or not evaluated thoroughly, for several reasons. First, at least for now, most of the tests published in Spain do not require consideration of these issues. Second, we wanted to avoid a drastic change from the original CET, in order to facilitate comparability with the results of previous evaluations, and in order to maintain a reasonable number of items to facilitate the reviewers' task.

It should be noted that all changes were made whilst generally keeping the structure, the sections and the way of scoring of the original CET (although in some cases further clarifications were made regarding the criterion of excellence and therefore the maximum score).

The initial proposal was reviewed in depth by the other authors of this paper and a new version of the CET-R was generated. This new version was reviewed by eleven qualified experts familiar with the CET, who are listed in Table 1. One of the most commonly suggested changes was to abandon the traditional classification of the types of validity that CET kept (and also the first version of CET-R) and to adopt the validity terminology of

the new standards of AERA, APA and NCME (2014). Therefore, it is this validity section which has undergone the biggest change compared with the original CET and the new EFPA model. In the terminology of the standards of the APA, AERA and NCME (1999, 2014) it is not the test that is validated but rather the interpretations or specific uses made of its scores. Therefore, instead of following the traditional classification of validity types of the APA from 1985 (AERA, APA, NCME, 1985), and differentiating between content, construct, and predictive validity, in CET-R, three sources of validity evidence are collected: evidence based on the content, evidence based on relationships with other variables (with another test that measures the same or a related construct, with a criterion that seeks to predict, etc.), and evidence based on the internal structure of the test (for example, evaluating the factor structure). In fact, the important thing is that evidence is gathered in the documentation and test manual to support the validity of using scores, regardless of whether talking about construct validity, or evidence of validity based on the internal structure of the test (formerly considered "construct validity"), for example. In fact, the updated EFPA model still uses the traditional classification. However, we believe that the update of the CET should incorporate the recommendations of the current international standards.

After making the relevant adjustments and modifications based on the suggestions of the experts, this new version was presented to the COP Test Commission, leading to the final version. This version, along with the completion instructions included in the questionnaire in order to increase clarity and standardization in the evaluation process, can be downloaded from the website of the Spanish Psychological Association (http: // www.cop.es), in the test commission section (or directly from http://cop.es/n).

Like the original model, the CET-R is divided into three sections. The first, focused on the technical description of the test, now has 28 items. It is virtually identical to the original except that some items include additional explanations and/or some answer options have been modified or added. In addition, two of the items of the original CET, items 1.20 and 1.21, referring to the scales used and transformed scores, respectively, are merged into one, and other items, such as the one concerning the presentation of the basic literature provided, have been eliminated.

The second section deals with the technical evaluation of the characteristics of the instrument. It includes 55 items, nine on general issues, one on item analysis, 20 on validity, 15 on reliability and 10 on scales and the interpretation of scores. Added to the initial items of the CET, concerning the quality of materials and documentation, the theoretical foundation, etc., is an item referring to the development of the items (when it is an original test, not an adapted one). It is also differentiated between the quality of the instructions for those who have to respond to the test, and for those who have to administer it and correct it, and an item is added that evaluates the quality of the references provided. The section on validity is the one with the most changes, as stated above. Along with the evidence based on the content, evidence based on the relationships with other variables is evaluated, differentiating between evidence based on relationships between test scores and other variables (convergent evidence, discriminant evidence, evidence based on differences between groups, etc.), and evidence based on relationships between test scores and a criterion (which would be the predictive validity in the original CET model). In addition, evidence is evaluated based on the internal structure, including at this point both the factor analysis and DIF analysis. Finally, an item is introduced which includes whether the manual reports the possible adaptations to be made in the administration of the test for the correct assessment of people with functional limitations or diversity. As for the section on reliability, in the evaluation of the coefficients of equivalence (parallel forms) an item is added on the evaluation of compliance with the assumptions of parallelism, and considering the coefficients of internal consistency, coefficients are added based on the factor analysis. Three questions are also included (two evaluative and one purely descriptive) concerning the quantification of the score precision using IRT, as well as two questions regarding the assessment of the inter-rater reliability. As for the section of norms and interpretation of scores, the norm-referenced interpretation includes one question on continuous norming, which allows us to obtain more accurate norms with smaller groups (e.g., Evers et al., 2010), and another question on updating the norms. In addition, four questions are included on the criterion-referenced interpretation of test scores specifically applicable to certain types of tests (e.g., educational or clinical). It should be noted that, for all sections in some items, clarifications are added about how to respond or the meaning of compliance with the criterion of excellence. Moreover, as in the original CET, open questions are included that enable us to justify the scores assigned to the closed items as well as other descriptive and evaluative questions that may be relevant.

Finally, in the third and final CET-R section, an overall assessment of the test is requested, as well as a summary of the first two sections which is reflected in a data sheet.

TABLE 1
LIST OF EXPERTS WHO PARTICIPATED IN THE REVIEW OF THE FIRST VERSION OF THE CET-R

Name	Affiliation
Constantino Arce	Universidad de Santiago de Compostela
Rosario Martínez-Arias	Universidad Complutense de Madrid
Roberto Colom	Universidad Autónoma de Madrid
Ana Delgado	Universidad de Salamanca
Eduardo Fonseca	Universidad de La Rioja
María Dolores Hidalgo	Universidad de Murcia
María José Navas	UNED
Julio Olea	Universidad Autónoma de Madrid
José Luis Padilla	Universidad de Granada
Pablo Santamaría	TEA Ediciones
Carme Viladrich	Universidad Autónoma de Barcelona

CONCLUSIONS

We believe that implementation of the CET model proposed by Prieto and Muñiz (2000) has had a positive impact in many areas in recent years. Firstly, it has provided test users with technical information about the quality of some of the tests available (almost 50 to date), to help them in their decision. But also, secondly, the application of the CET has helped to improve the processes of construction and publishing the tests. Over the various assessments that have been carried out, we have observed that, increasingly, the test manuals explicitly include most of the CET evaluation criteria, and they include detailed information on the processes of construction and standardization of the test, the psychometric quality of its scores and the appropriate and inappropriate uses of the test. Finally, we are aware that the CET is having an impact on the training of future psychologists, since teachers of psychometrics often use this model in their classes, guiding students in a practical way in the basics of evaluating the psychometric and technical quality of tests.

Recognizing this is not contradictory to accepting that, after more than 15 years since the publication of the CET, the concepts of reliability and validity have been enriched, and the scientific and professional requirements of the test have been adapted to new needs (De Boeck & Elosua, in press). Therefore, a review of the CET model that would incorporate the progress made was necessary in order to incorporate improvements in the use of tests by psychologists and educators, and indirectly, to further improve the processes of constructing and publishing tests in our country. This review has materialized in the CET-R, to be used in the fifth edition of test evaluations, driven by the COP, which has been launched recently.

The publication of the results of the test evaluations is one of the informative strategies the COP follows in order to improve the use of tests and thus the professional practice of psychologists. But it is not the only one. The COP, along with the EFPA and the ITC (International Test Commission), of which it is a member, carries out varied activities and projects in order to improve the use of tests. The various activities and projects are part of two complementary strategies: one which is more restrictive and the other informative (for more detailed information see Muñiz & Bartram, 2007; Muñiz & Fernández-Hermida, 2010, and Muñiz, 2012). The restrictive strategy comprises the totality of activities carried out to limit the use of tests to professionals who are actually qualified to do so. The informative strategy brings together initiatives to disseminate information on the practice of tests in order to reduce the likelihood of misuse of tests. In this regard, ethical and professional codes have been developed (e.g., EFPA, 2005; Fernández-Ballesteros et al, 2001) and guidelines on the use of tests, including the technical standards of the AERA, APA and NCME (2014) as well as numerous guidelines developed by the ITC, have been proposed: the general guidelines for the use of tests, (ITC, 2001), the guidelines for the translation and adaptation of tests from one culture to another (Hambleton, Merenda & Spielberger, 2005; Muñiz, Elosua & Hambleton, 2013), the guidelines on the use of computerized tests, the professional guidelines on the selection of tests and how to proceed when tests become obsolete, the guidelines on the security of tests, on the quality control of tests, and the use of tests in research. The most important of the last three is reflected in the work of Muñiz, Hernández and Ponsoda (2015). All are available on the website of the ITC and many of them have been translated into Spanish and are accessible through the website of the Spanish Psychological Association (http://www.cop.es) in the section of the test commission. One final informative strategy that deserves attention is the IS0-10667 standard, which regulates the whole process of assessing people in work contexts. For a more detailed review of all of the actions taken in Spain to improve the use of tests, please see Elosua and Muñiz (2013).

The evaluation of the tests published in Spain is one of many actions. But, as Elosua and Geisinger (2016) indicated, for this action to be really useful, it requires continuous improvement work, both procedural as well as formal and substantive. And the CET-R is proposed with this improvement objective in mind. The ultimate goal is clear: for psychologists to have proven and reliable information that will help them make a better selection and use of the available tests. All of this will impact on improving professional practice and its prestige.

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *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. (2002). *Review model for the description and evaluation of psychological tests*. Brussels: European Federation of Psychologists' Associations (EFPA).

Bartram, D., Lindley, P. A. & Foster, J. M. (1990). *A review of psychometric tests for assessment in vocational training*. Sheffield, UK: The Training Agency.

Bartram, D., Lindley, P.A. & Foster, J.M. (1992 a). *A review of Psychometric tests for assessment in vocational training*. Leicester: BPS Books.

Bartram, D., Lindley, P.A. & Marshall, L. (1992 b). *Update to the review of psychometric tests for assessment in vocational training*. Leicester: BPS Books.

Bartram, D., Anderson, N., Kellett, D., Lindley, P.A. &

Robertson, I. (Eds.). (1995). *Review of Personality Assessment Instruments (Level B) for use in occupational settings*. Leicester: BPS Bools.

Bartram, D., Burke, E., Kandola, R., Lindley, P., Marshall, L. & Rasch, P. (Eds.). (1997). *Review of Tests of Ability and Aptitude (Level A) for use in occupational settings*. Leicester: BPS Books.

Buros, O. K. (1938). *The 1938 Mental Measurements Yearbook*. New Brunswick, NJ: Rutgers University Press.

Carlson, J. F., Geisinger, K. F. & Jonson, J. L. (Eds.) (2014). *The nineteenth mental measurements yearbook*. Lincoln, NE: Buros Center for Testing.

Carlson, J. F. & Gonzalez, S. E. (2015). Using Pruebas Publicadas en Español to enhance test selection. In K.F. Geisinger (Ed.), *Psychological testing of Hispanics (2nd ed.): Clininal and intelectual issues (pp. 11-27)*. Washington, DC: American Psychological Association.

De Boeck, P. & Elosua, P. (in press). Reliability and Validity: History, Notions, Methods, Discussion. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC intemational handbook of testing and assessment*. Oxford University Press.

European Federation of Professional Psychologists' Associations (2005). *Meta-Code of ethics*. Brussels: Author (http://www.efpa.eu)

Elosua, P. (2012). Tests publicados en España: Usos, costumbres y asignaturas pendientes [Tests Published in Spain: Uses, Customs and Pending Matters]. *Papeles del Psicólogo, 33,* 12-21.

Elosua, P. & Geisinger, K (2016) Cuarta evaluación de tests editados en España: forma y fondo [Fourth Review of Tests Published in Spain: Form and Content]. *Papeles del Psicólogo, 37, 82-88*.

Elosua, P. & Muñiz, J. (2013). Proyectos españoles para una mejora en el uso de los tests [Spanish projects for improving the use of tests]. *Revista Latinoamericana de Ciencia Psicológica*, 5, 139-143.

Evers, A., Braak, M., Frima, R. & van Vliet-Mulder, J. C. (2009). *Documentatie van Tests en Testresearch in Nederland [Test documentation and research on tests in the Netherlands]*. Amsterdam: Boom test uitgevers.

Evers, A., Muñiz, J., Hagemeister, C., HstmwLingen, A., Lindley, P., Sjóberg, A. & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema, 25,* 283-291.

Evers, A., Sijtsma, K., Lucassen, W. L & Meijer, R. R. (2010) The Dutch Review Process or evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing, 10,* 295-317.

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: LEA.

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, 36,* 1-8.

ISO (2011). *Procedures and methods to assess people in work and organizational settings (part 1 and 2)*. Geneva: ISO [Spanish version: Prestación de servicios de evaluación: procedimientos y métodos para la evaluación de personas en entornas laborales y organizacionales (partes 1 and 2). Madrid: AENOR, 2013].

Lindley, P.A. (2009, Julio). Using EFPA criteria as a common standard to review tests and instruments in different countries. In D. Bartram (Chair), *National approaches to test quality assurance*. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

Moosbrugger, H., Kelava, A., Hagemeister, C., Kersting, M., Long, F., Reimann, G., et al. (2009, July). The German Test Review System (TBS-TK) and first experiences. In D. Bartram (Chair), *National approaches to test quality assurance*. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

Muñiz, J. (2012). Perspectivas actuales y retos futuros de la evaluación psicológica [Current perspectives and future challenges of psychological assessment]. In C. Zúñiga (ed.), *Psicología, sociedad y equidad [*Psychology, Society and Equity*]*. Santiago de Chile: Universidad de Chile.

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

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,* 108-121.

Muñiz, J., Elosua, P., & Hambleton, R.K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición [Guidelines for the translation and adaptation of tests: Second edition]. *Psicothema, 25,* 151-157.

Muñiz, J., Fernández-Hermida, J. R., Fonseca-Pedrero, E., Campillo-Álvarez, A., & Peña-Suárez, E. (2011). Evaluación de tests editados en España [Review of Tests Published in Spain]. *Papeles del Psicólogo, 32,* 113-128.

Nielsen, S. L. (2009, Julio). Test certification through DNV in Norway. In D. Bartram (Chair*), National approaches to test quality assurance*. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

NIP (1969). *Documentatie van Tests en Testresearch in Nederland [Test documentation and research on tests in the Netherlands]*. Amsterdam: Nederlands Instituut van Psychologen.

Ponsoda, V. & Hontangas, P. (2013). Second evaluation of tests published in Spain. *Papeles del Psicólogo, 34,* 82-90.

Prieto, G. & Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España [A model for evaluating the quality of the tests used in Spain]. *Papeles del Psicólogo, 77,* 65-72.