Papeles del Psicólogo, 2015. Vol. 36(1), pp. 1-8
http://www.papelesdelpsicologo.es

THIRD EVALUATION OF TESTS PUBLISHED IN SPAIN

Ana Hernández, Inés Tomás, Adoración Ferreres and Susana Lloret
Metodología de las Ciencias del Comportamiento [Methodology of Behavioural Sciences] and IDOCAL.*
Universitat de València

In order to provide technical information on the tests published in Spain, and continuing with the evaluation of some of the tests that are most frequently used in this country, this paper describes the results of the third evaluation of psychological and educational tests carried out by the Test Commission of the Spanish Psychological Association. In total, 11 tests were evaluated. As in the previous evaluations, each test was evaluated by two experts who responded to the Questionnaire for the Assessment of Tests – CET- (Prieto & Muñiz, 2000). However, considering that the European Federation of Psychological Associations has recently revised the European model -which was partially based on the CET model-, for this third evaluation we have introduced a number of changes that are described in this paper. In addition, for each test, results are presented regarding the quality of the documentation and materials, the theoretical foundation, the Spanish adaptation, the item analysis, the coverage of the validation studies, the reliability, and the norms. Both quantitative and qualitative information is provided, highlighting the main strengths and weaknesses of the test. Finally, we also present a number of recommendations and suggestions for future test evaluations.
Key words: *Tests, Test use, Test evaluation, Psychometric properties.*

Con el fin de proporcionar información técnica sobre los tests editados en España, y siguiendo el proceso de evaluación de algunos de los tests más empleados en nuestro país, el presente artículo presenta los resultados de la tercera evaluación de tests llevada a cabo desde la comisión de tests del Colegio Oficial de Psicólogos. En concreto se han evaluado un total de 11 tests. Como en las dos evaluaciones previas, los tests han sido evaluados por dos expertos mediante el Cuestionario para la Evaluación de los Tests (CET) propuesto por Prieto y Muñiz (2000), si bien en la presente edición, y con motivo de la reciente revisión del modelo de evaluación elaborado por la Federación Europea de Asociaciones de Psicólogos Profesionales, se han aplicado algunas modificaciones que son pertinentemente señaladas. Para cada test se presentan los resultados sobre la calidad de los materiales y su documentación, la fundamentación teórica, la adaptación española (si procede), el análisis de ítems, las evidencias de validez recogida, la fiabilidad de sus puntuaciones, y la calidad de los baremos. Se ofrecen tanto resultados cuantitativos sobre estos aspectos como información cualitativa que resalta los puntos fuertes del test y los aspectos a mejorar. Por otra parte también se presenta información sobre el proceso de revisión y algunas cuestiones a mejorar para futuras ediciones.
Palabras clave: *Tests, Uso de los tests, Evaluación de tests, Propiedades psicométricas.*

Tests are a basic tool in psychological evaluation, and they help the practitioner to make decisions that can have major consequences for individuals. It is therefore necessary to ensure that the psychometric properties of the tests are appropriate and that they are used by competent professionals. One of the informative strategies being carried out by the Spanish Psychological Association (or College of Psychologists, COP) to achieve both of these objectives and to enhance the use of the tests consists of providing reliable information on the theoretical, practical and psychometric characteristics of the tests, which will help practitioners to make appropriate decisions and to use the tests correctly. In fact, psychologists have long been demanding this kind of technical information (see Muñiz et al., 2001; Muñiz & Fernández-Hermida, 2010).

In this context, in 2010, the Test Commission of the COP, following the lead of other countries such as the Netherlands and the United Kingdom, implemented the process of evaluation of the tests published in Spain. Specifically, the evaluation was performed by the Test Evaluation Questionnaire (CET in Spanish) (Prieto &

Correspondence: Ana Hernández Baeza. *Departamento de Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Universitat de València. Av. Blasco Ibáñez, 21. 46010 Valencia. España. E-mail: Ana.Hernandez@uv.es*

**Instituto Universitario de Investigación en Psicología de los Recursos Humanos, del Desarrollo organizacional y de la Calidad de Vida Laboral [University Research Institute of Human Resources Psychology, Organizational Development and Quality of Working Life]*

Muñiz, 2000) which inspired, together with other European models, the evaluation model proposed by the testing committee of the European Federation of Professional Psychologists Associations (EFPA) (Evers et al., 2013). One of the most important features of the CET is that it enables the provision of both quantitative and qualitative information on the psychometric quality of the test evaluated, as well as on the quality of the materials and documentation.

Following this model, in 2011 the results were published of the first evaluation of tests published in Spain (Muñiz, Fernández-Hermida, Fonseca-Pedrero, Campillos-Alvarez & Peña-Suarez, 2011), in which a total of 10 tests were assessed (namely, the WAIS-III, WISC-IV, MCMI-III, MMPI-2-RF, 16PF-5 PROLEC-R, EFAI, NEO-PI-R, EVALUA, and IFG tests). Two years later the results of the second evaluation were published (Ponsoda & Hontangas, 2013), in which a total of 12 tests were assessed (namely, the BAI, BAS-II, BDI-II, CEAM, CompeTEA, EPV-R, ESCOLA, ESPERI, Merill-Palmer.R, PAI, RIASRSIT, and WNV tests). Detailed reports of the tests reviewed are available on the website of the COP (in the Test Commission section, and in the sub-section Test Evaluation in Spain). Also, the summary of the results and the process carried out for both evaluations can be found in the work of Muñiz et al. (2011) and Ponsoda and Hontangas (2013), published in this journal.

TABLE 1
LIST OF TESTS EVALUATED

BCSE	Test Breve para la evaluación del estado cognitivo [Brief Cognitive Status Exam]
BECOLE	Batería de Evaluación Cognitiva de la Lectura y la Escritura [Battery of Cognitive Evaluation on Reading and Writing]
BOHEM-3	Test Boehm de conceptos básicos [Boehm Test of Basic Concepts]
BOHEM-3 PRESCOLAR	Test Boehm de conceptos básicos - 3 Preescolar [Boehm Test of Basic Concepts – Preschool 3]
CESQT	Cuestionario para la Evaluación del Síndrome de Quemarse por el Trabajo [Questionnaire for the Evaluation of Occupational Burnout Syndrome]
ECLE	Evaluación de la comprensión lectora [Evaluation of Reading Comprehension]
ESQUIZO-Q	Cuestionario Oviedo para la Evaluación de la Esquizotipia [Oviedo Questionnaire for Schizotypy Assessment]
IECI	Inventario de Estrés Cotidiano Infantil [Children's Daily Stress Inventory]
SOC	Escala de Dificultades de Socialización de Cantoblanco [Cantoblanco Scale for the assessment of Socialization Difficulties]
TRAUMA	Test de Resistencia al Trauma [Resistance to Trauma Test]
WAIS-IV	Escala de inteligencia de Wechsler para adultos-IV [Wechsler Adult Intelligence Scale – IV]

In this article we present the results of the third evaluation of tests, in which 11 tests are rated (specifically, the BCSE, BECOLE, BOHEM-3, BOHEM-3 Preescolar [Pre-school], CESQT, ECLE, ESQUIZO-Q, IECI, SOC, TRAUMA and WAIS-IV) (see Table 1). Firstly, prior to presenting the results, we report some modifications that have been made both to the process and to the CET evaluation questionnaire itself. While the review process was similar to that used in the previous editions in general, these small changes must be considered when comparing the results with those obtained in the previous evaluations. Secondly, the overall results of the evaluated tests are presented, highlighting their strengths and the areas that can be improved. Finally, based on the results and our experience in the review process, some recommendations are given that may be useful for future assessments.

THE EVALUATION PROCESS

In this third evaluation, the overall process is basically consistent with the one followed in the previous evaluations, although there are some differences that should be considered. Firstly, whereas in the previous evaluations the Test Commission selected the tests to be reviewed, on this occasion it was the publishers who selected the tests that they wished to submit for review, and the proposal was accepted unanimously by the Test Commission. As noted above, a total of 11 tests were selected (three from the editorial EOS, four from Pearson and four from TEA). Secondly, and as in previous assessments, the coordinating team appointed by the Test Commission (led by the first author of this article), selected a group of reviewers, in order to allocate two per test: one reviewer with a more technical psychometric profile and another with a more theoretical profile, an expert in the substantive aspects of the variable measured by the test. Efforts were made to ensure that the reviewers did not have a direct relationship with the authors of the tests, and that they had no conflict of interest. (In fact, in the letter of invitation to participate in the process, they were told that they should not participate if they doubted they were able to carry out an objective evaluation). In some cases, the reviewers that were initially selected declined to participate in the process for various justified reasons, so

it was necessary to choose another reviewer. The final list of participating reviewers is shown in Table 2. We sincerely thank all of the reviewers for their positive response and involvement in the process.

The editors gave three complete copies of each test to the COP. The COP sent one to each reviewer and the third to the coordinator. In the past, the tests were only sent to the reviewers, however, this time, following the recommendation of Ponsoda and Hontangas (2013), a third set was sent to the coordinator in order to facilitate the task. In addition to the test, the reviewers were paid a token amount of 50 euros (which some chose to decline). As on previous occasions, the task of the reviewers was to implement the CET and evaluate it using the assigned test. However, considering that the test evaluation model developed by the testing committee of the EFPA was recently reviewed, the Test Commission inspected this new model and its changes and decided to include some new items in the CET model, and/or modify some specific issues. The EFPA revised model is available in English from the website http://www.efpa.eu/professional-development, (see the section dedicated to Assessment). Also, in the article by Evers et al., (2013), published in the journal Psicothema, an English summary is available of the main changes implemented. In the case of the CET, the main changes made in this edition are described below.

With regards to the section concerning the general description of the test, some general areas of content on the variable measured in the test were added, as well as some new areas of application (items 1.11 and 1.12 of the original CET published by Prieto and Muñiz (2000)). Also, for the item referring to the transformation of scores (item 1.21 of the original CET), the clarification was added that the normalised transformation referred to the scores obtained by normalisation applied using the standard normal distribution, while the non-normalised one referred to standardised scores which were obtained using linear transformations. Therefore, when only percentiles were given, neither of the two options were relevant in this new version of the CET, and it had to be indicated that it was not applicable. The percentile scores are now provided in detail in describing the type of scale (item 1.22 of the original CET), differentiating between the various types of percentiles (percentiles, quintiles and deciles) as well as the standardised scores and their derivatives (decatypes, stanines, T, etc). Finally, added to the item referring to the documentation submitted by the publisher was the option "Complementary technical information and updates" (item 1.26 of the original CET).

As regards the evaluation of the properties of the test, a number of modifications were also performed. Firstly, in evaluating the evidence of the construct validity (item 2.10.2.1 of the original CET), other options were added to the existing ones: correlations with other tests, and analysis of invariance/differential item functioning (DIF); also, added to the existing option of experimental design was quasi-experimental design. In addition, 3 items were added to explicitly evaluate the results of the differences between groups (these could be natural or experimental), the results of the analysis of the multitrait-multimethod matrix, and the results of the factor analysis. In terms of the reliability section, in item 2.11.1 of the original CET, referring to the information provided on reliability, the option of "Quantifying the error by IRT (information function or others)" was added to the existing options. Also added were an item to assess the adequacy of the sample size when quantifying the error by IRT and another item to report the internal consistency coefficients presented. Finally, with respect to the section on scales, an item was included to evaluate the updating of these scales.

TABLE 2
REVIEWERS WHO EVALUATED THE TESTS

Revisor	Afiliación
Francisco José Abad García	Universidad Autónoma de Madrid
Amelia Catalán Borja	Centre de Psicologia Clinica i Formativa, Valencia
Paula Elosua Oliden	Universidad del País Vasco
Antonio M. Ferrer Manchón	Universidad de Valencia
Adoración Ferreres Traver	Universidad de Valencia
Eduardo Fonseca-Pedrero	Universidad de La Rioja
Maite Garaigordobil Landazabal	Universidad del País Vasco
José Manuel García Montes	Universidad de Almería
Luis F. García Rodríguez	Universidad Autónoma de Madrid
Remedios González Barrón	Universidad de Valencia
Giorgina Guilera Ferré	Universidad de Barcelona
Mª Dolores Hidalgo Montesinos	Universidad de Murcia
Susana Lloret Segura	Universidad de Valencia
Sonia Mariscal Altares	UNED
Isabel Martínez Sánchez	Universidad de Castilla-La Mancha
José Carlos Núñez Pérez	Universidad de Oviedo
Julio Olea Díaz	Universidad Autónoma de Madrid
José Luis Padilla García	Universidad de Granada
Herminia Peraita Adrados	UNED
Jesús Pérez Hornero	Hospital de Conxo, Santiago de Compostela
Ingeborg Porcar Becker	Universidad Autónoma de Barcelona
Patricia Recio Saboya	UNED

Moreover, taking into account the recommendations made in the previous evaluations (see Muñiz et al., 2011 and Ponsoda & Hontangas, 2013), in this third evaluation, the Test Commission decided to send general instructions on how to complete the CET, in order to reduce ambiguities and to standardise the process more. These general instructions are presented in Appendix 1. In addition, a glossary of psychometric terms was provided to serve as a guide and reminder, especially aimed at the reviewers with a more theoretical profile.

After receiving the evaluations from the reviewers of each test, the coordinating team integrated the two reviews, generating a report for each test. When there was disagreement between the reviewers, the coordinating team conducted an independent assessment using the materials provided. Also, depending on the characteristic assessed, a differential weight was given to the two evaluations based on the profile of the reviewers. With this, the final score was awarded and the final evaluation was performed. As with the previous evaluations, the reports were then sent to the editors, so that both they and the authors had the opportunity to clarify and refine some of the comments of the reviewers and, ultimately, to present their point of view. These clarifications and insights were integrated into the final report, modifying the evaluations when it was considered to be justified. We would like to emphasize the professionalism of the authors and editors in responding to the review and thank them for their commitment to introduce some of the suggested improvements for future prints and editions of the manuals.

RESULTS OF THE EVALUATION AND CONCLUSIONS

The main results obtained for the 11 tests evaluated can be found in Table 3. Note that in no case were there parallel forms of the tests evaluated, so this aspect (reliability: equivalence) is not included in the table. Considering that the items are rated using a 5-point scale (1 = inadequate, 2 = adequate but with some shortcomings, 3 = adequate, 4 = good, 5 = excellent), and that mean values equal to or greater than 2.5 are considered adequate, it is observed that, for the vast majority of the characteristics evaluated, the vast majority of tests are, at least adequate, and in many cases good (3.5 onwards) or even excellent (4.5 onwards). Through the various tests, the average ratings for each of the characteristics evaluated were in no case below the mid-point 3 (adequate), the highest average in this evaluation being 4.1 (good).

Thus, it can be concluded that the tests that were evaluated are of reasonable quality. Logically there is some variability, and in general all of the tests have

TABLE 3
SUMMARY OF THE RATINGS OF THE TESTS EVALUATED

	Tests											Evaluation Means		
Characteristics	BECOLE	TRauma	ECLE	CESQT	ESQUIZO-Q	IECI	SOC	BCSE	Boehm-3	Bohem-3 Preescolar	WAIS-IV	3rd	2nd	1st
Quality of the materials and documentation	3	4	3	5	4.5	3.5	4.5	4.5	4.5	4	5	4.1	4.3	4.4
Theoretical basis	3.5	2	3	4	4.5	3.5	4.5	2.5	4	3	5	3.6	4.0	4.2
Spanish adaptation	-	-	-	-	-	-	-	3.5	4	3	5	3.9	4.3	4.3
Item analysis	3.5	3	-	4	4.5	4	4	-	-	-	3	3.7	3.9	3.6
Content validity	-	3	2	4	3.5	4.5	4.5	2.5	3.5	4.5	3.5	3.6	3.5	4.3
Construct validity	3.5	3.5	3.5	4	4	4	4	3	4	3	5	3.8	4.1	3.6
Analysis of bias	-	-	-	-	4.5	-	4	-	-	-	-	4.3	5	-
Predictive validity	3.5	-	3.5	3.5	-	2.5	4.5	-	-	-	4	3.6	3.6	3.6
Reliability: internal consistence	4	4	4	3	3.5	3	3.5	-	3	3.5	4.5	3.6	4.0	4.5
Reliability: stability	-	-	-	3	-	3	1	3	4	4	3	3.0	4.2	3.8
Norms	4	4	4.5	4	4.5	4	3.5	4	3	3	5	4.0	3.4	3.5

Notes: No test had parallel forms, so this way of evaluating the reliability was not included in the table.
The scores in the table are given on a scale whose 5 values are: 1= inadequate, 2=adequate but with some shortcomings, 3= adequate, 4= good, 5= excellent. Where there is a dash (-) it means that the information was either not provided or not necessary.

strengths and other aspects that could be improved. Specifically, the aspects that most often require attention in the various tests are the theoretical foundation presented, the content validity and the estimation of reliability as stability. The latter aspect is the one with the lowest average, having a score of 3. In the full assessments, available on the website of the COP: www.cop.es, in the Test Commission section, information is given justifying the scores awarded and explaining how each aspect could be improved, either with additional information, since in some cases the information is not complete or clear enough, or by conducting additional studies or increasing the size of the samples used. Also noteworthy are the lack of information and analysis related to the item analysis, the predictive validity, the reliability and stability, and the analysis of bias or DIF. All of this information should be added gradually and progressively. If there are cases where it is not relevant to obtain information on any of these aspects (for example referring to variables that are expected to change over short periods of time) this could be made explicit in the manual.

Table 3 also shows the means of the ratings awarded for the different characteristics considered in the various evaluations. However it should be remembered that these means are not directly comparable in many cases, since, in this third evaluation, some more detailed instructions have been added on how to evaluate certain aspects, and a number of items have been added (for reliability, construct validity and scales), thus evaluating new aspects. We should emphasize that the number of tests that analyse bias and differential item functioning has gradually increased. Although DIF studies are only performed for two of the 11 tests, in the previous evaluations published in 2011 and 2013 there were none and one, respectively. In this latest revision, there has also been a reduction in the proportion of adapted tests selected for evaluation, increasing the number of original tests that are applicable to the Spanish context. It is also worth noting the introduction of the use of Item Response Theory (IRT) models in some of the tests evaluated. Finally, it is interesting to note that in some cases and for the most recent tests, the manuals follow the sections of the CET model in considerable detail, which leads us to believe that this process of test evaluation in Spain, in sending a clear message to the editorials on the quality criteria that are required, may be having an impact with regards to improving the information presented in the manuals and, consequently, their quality.

LOOKING TO THE FUTURE

To date, a total of 33 tests have been assessed using the CET model in Spain (Prieto & Muñiz, 2000). The objective is to continue with the review process until, ideally, almost all of the tests have been reviewed, as occurs in other countries such as the Netherlands (see Evers, Sijtsma, Lucassen, & Meijer, 2010). However, the technological and psychometric advances of recent years, and the experience gained in the review process over these years, recommend both a review of the model and the introduction of some changes in the review process, although some of the latter may be debatable.

With regards to the CET model, it was noted earlier that in this third test evaluation some items have been modified or incorporated based on the recent revision of the test evaluation model proposed by the testing committee of the EFPA. However, it is still necessary to conduct a more thorough review of the European model, and to gradually incorporate changes that enable us to assess in detail, among other things, the computerized administration of tests, remote evaluation via the Internet, the quality of automated reports, the application of Item Response Theory, criterion-referenced tests and continuous norming. In this revised CET, the suggestions made by Muñiz et al. (2011) to add evaluative questions referring to the bibliography, and to the process of translation/adaptation of the test should also be taken into account (see also Elosúa, 2012). We hope to have a revised CET model soon which will enable us to assess these issues in forthcoming evaluations.

As regards the evaluation process itself, and related to the application of the model, we believe it would be appropriate to differentiate between information that is not provided because it is not applicable to the test (e.g., information on the adaptation process when it is a test that is constructed and analysed using Spanish samples), and information that is not provided, despite the fact that it would be informative and relevant in assessing the test quality. This lack of differentiation may send the wrong message to the editorials regarding the importance of providing all of the available information, even when the results are not entirely suitable. The test, in its evaluation, may have a better final score if this information is not presented, which would go against the spirit of these

evaluations completely. However, we must remember two things. First, the professional psychologist can also consider this differentiation when making decisions (information not being provided because it is unnecessary is not the same as it not being provided when it would be relevant). Furthermore, the process of validation and evaluation of the metric quality of the test is a continuous process that is not completely closed at the time the manual is published. New editions of the manuals may add new studies and information. We agree with Ponsoda and Hontangas (2013) that this could increase the price of the manual, jeopardizing its commercialization. As a possible alternative, this additional information could be published by the editorials on their websites as they carry out further studies.

Still on the process of applying the CET, it is worth mentioning that, despite the additional instructions provided to the reviewers (see Appendix 1), there still seem to be some ambiguities and unclear issues when applying the model. Here we summarize some of the issues that we believe could be improved. First, in the section on Item Analysis, an evaluation of the psychometric information of the items is what is requested and not whether they are more or less correct in terms of grammar and the language used. But this is not always understood by the reviewers. Also, at times, even though the item analysis does not appear as a separate section in the manual, an item analysis is in fact performed, which goes unnoticed by some reviewers. Therefore it is important to note this in the CET, so the reviewers do not mark the "no information is provided in the documentation" option when it appears in other sections such as those of validity analysis or reliability and internal consistency. Second, another issue that we feel is not sufficiently clear is the item "the quality of the tests used as criteria or markers" and what it refers to. It seems that sometimes it is interpreted as the appropriateness of the selection of the criterion based on the theories of the construct, and others, as the metric quality of the tests used to measure the criterion. Also, instead of appearing in the section of construct validity, it might be more appropriate for this item to appear in the section of criterion validity (concurrent and predictive). Third, and concerning the analysis of sensitivity and specificity which enable the evaluation of the diagnostic accuracy of the test, at times the results are presented as differences between groups (under the heading of construct validity) and at other times as predictive validity. Explicit information should be provided on which section should include this analysis. Fourth, and still on the subject of construct validity, in evaluating the median of the correlations with other similar tests the task is not entirely clear. Since the manuals often include numerous correlations with different tests and criteria within the same table, the correlations that are to be considered should be clarified or examples provided. Finally, the section "Basic bibliography regarding the test provided in the documentation" is ambiguous. While some reviewers assess its adequacy and whether it is up to date, others merely mention some of the references provided. Perhaps it would be good if these issues (and others raised in the previous assessments) were to be introduced as clarifications in the corresponding section of the CET to facilitate a more standardised application. However, it is also the responsibility of the editorials to present the information as clearly as possible, with clear sections that enable the localisation of information and presenting clear validation hypotheses.

As regards the reviewers, it is debatable whether the evaluations of the different profiles selected for each test (one more technical-psychometric and the other an expert in the construct of concern) should be weighted differentially depending on the specific aspect being assessed in the CET. In any case, we believe it is crucial that the coordinating team continues to have a complete set of the test in order to perform a better integration of the evaluations of the reviewers taking their profiles into consideration, and also to respond to feedback from the publishers regarding the provisional report produced after the review. Another aspect to consider is that the reviewers with a more psychometric profile may sometimes be too demanding, and they may ask for information and analysis that the professional psychologist who will use the test would probably not understand. We believe that these analyses and information are important and can provide crucial evidence about the psychometric quality of the test, so a possible solution would be, again, to publish the most complex studies with all the technical information on the website of the editorial, thus increasing the transparency. The reviewers, when taking into account this information, should make evaluative comments aimed at the practitioners who may not have knowledge about the latest psychometric advances.

However, and in relation to the last point mentioned, the psychometric expertise of the professionals, we must not forget that it is the responsibility of every good test user to continue to educate themselves and learn about the new psychometric advances. In fact, some countries have accreditation systems that aim to guarantee that users have sufficient knowledge tests and skills for the proper use of tests, and some countries, such as the UK, have even implemented a European accreditation system that adds value to this accreditation. It should be noted that the COP is part of the European accreditation committee of test users (the TUAC, Test User Accreditation Committee, represented by the first author) although, to date, no action has been taken to implement this process. If we consider that psychologists who responded to the latest survey of attitudes toward tests (Muñiz & Fernández-Hermida, 2010) admitted that the training received in the psychology degree may not be enough for the correct use of most tests, and considering that psychometric knowledge progresses in such a way that the distance with applied practice is today greater than ever (Elosúa, 2012), we believe it would be interesting to open this route of accreditation in Spain, which could motivate professionals to retrain and to update their knowledge on the advances in psychometrics.

In summary, we are still far from having evaluated all of the tests published in Spain. However, we believe that we are on the right track. Although there are still many areas for improvement, we believe that the evaluation process initiated by the COP a few years ago works well overall and is gradually having an impact on improving the tests and manuals. It is also important to note the results of the tests evaluated throughout the various editions, which are generally satisfactory. In fact some of them have highly positive evaluations. Perhaps it would be good if the Test Commission were to grant "quality awards" to the best tests, as recognition and as an incentive towards excellence.

ACKNOWLEDGMENTS

We would like to thank the members of the COP Test Committee who have participated in the meetings and in the process, for their collaboration and assistance. Specifically José Ramón Fernández-Hermida, Miguel Martínez, Milagros Antón, Pablo Santamaría, Viviana Gutman, Frederique Vallar and especially José Muñiz and Vicente Ponsoda, the coordinators of the previous evaluations, for their continued support. We also reiterate our thanks to the reviewers for their willingness and professionalism.

NOTES

The first author wishes to clarify that, although her name is the same as one of the authors of the adaptations of some of the tests evaluated, it is not the same person and she has no professional relationship with the editorial.

REFERENCES

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.

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

Evers, A., 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*, 283-291

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. (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., 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.

Ponsoda, V. & Hontangas, P. (2013). Segunda evaluación de tests editados en España [Second evaluation of tests published in Spain]. *Papeles del Psicólogo, 24,* 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,* 65-71.

APPENDIX 1
OBSERVATIONS FOR REVIEWERS TO BEAR IN MIND WHEN RESPONDING TO THE CET

The CET has been applied in two review processes and in both we have seen that its completion involves some difficulties. The problems encountered are described in articles that report on the two reviews (published in the journal *Papeles del Psicólogo* in 2011 and 2013, volumes 32 (2) 34 (2) respectively).

The following lines are intended to give some hints as to how to resolve these issues. If you have questions, do not hesitate to ask me.

1) The CET is a test, and as such, its questions and options should not be changed. In past editions, there have been occasions where the reviewer, not finding the answer option that he or she was looking for, modified one of the existing options. You must, however, respond using the options that CET offers. If you do not find the option that you are looking for, choose the one that is most similar. The CET asks for justifications of only a few responses, not the majority. However, when it requests, for example, "comments on validity in general," the main justifications for the scores given to all of the questions on validity are expected.

   In principle, no question should be left unanswered. For a particular type of test, the CET may not be entirely appropriate. For example, the strategy of interpretation of the scores considered in the CET are the scales, though some tests (e.g., clinical scales) use other alternative interpretation strategies that are not explicitly listed in the CET. In this case, the questions for "scales" can be left blank, responding instead in "comments on scales" on the alternative procedures of interpretation and their quality, providing the necessary justifications.
2) In adapted tests, another important issue is how much weight to give to the studies carried out with the original test and the studies carried out in the adaptation process. Our position on this matter is that all studies submitted should be considered, although it seems reasonable to give more importance and weight in the evaluation to those that contribute to the adaptation process.
3) In the commercialised tests it is expected that the reviewer will carry out the review based on the documentation that is given to him or her. In the past it has happened that the manual omitted some information that the reviewer may consider important in evaluating the quality of the test. In this case, it is appropriate for the reviewer to ask the coordinator for this information, who, in turn, will ask the publisher for the required information and will see if it is possible to obtain it, and under what conditions.
4) If a battery or more than one test (e.g., normal and short test) is to be reviewed, two strategies can be followed. The one that the CET indicates is to fill in as many CETs as the number of tests that have to be revised. A less costly strategy which is also possible, if it makes sense, would be to only use one CET, noting where appropriate the different results obtained by the different tests.
5) In the first two reviews we have noticed that we do not all attach the same meaning to the psychometric terms used in the CET. Some misleading terms are discussed below. When the CET asks about the quality of items in the section of Item Analysis, what is required is an assessment of the psychometric information that the manual provides for the items, not whether upon reading them we deem that they are well or poorly worded, for example. A similar thing happens with questions about content validity. In fact, the aim is to find out what confirmation is provided regarding whether the test evaluates the relevant parts of the construct of interest.