Papeles del Psicólogo / Psychologist Papers, 2019. Vol. 40(1), pp. 21-30
https://doi.org/10.23923/pap.psicol2019.2886
http://www.papelesdelpsicologo.es
http://www.psychologistpapers.com

SIXTH REVIEW OF TESTS PUBLISHED IN SPAIN: RESULTS AND IMPACT OF THE MODEL ON LECTURERS AND PUBLISHERS

María Dolores Hidalgo[1] and Ana Hernández[2]
[1]*Universidad de Murcia.* [2]*Universitat de València*

El cuestionario para la Evaluación de los Tests (CET; Prieto y Muñiz, 2000) y su revisión (CET-R; Hernández et al., 2016) se han venido aplicando sistemáticamente desde 2010, impulsado por la Comisión de tests del Consejo General de la Psicología del Colegio Oficial de Psicólogos. El objetivo es proporcionar información contrastada sobre la calidad de las pruebas a los profesionales, con el fin de mejorar el uso de los tests. El presente trabajo tiene un doble objetivo. El primero, presentar los resultados de la sexta evaluación de tests psicológicos y educativos, en la que se han revisado un total de 10 tests. El segundo, evaluar el impacto que la aplicación del CET/CET-R ha tenido durante estos años en dos agentes cruciales: las editoriales de tests, y los profesores universitarios encargados de formar a los futuros profesionales usuarios de tests. Los resultados de la sexta evaluación, así como los resultados de la encuesta para evaluar el impacto del CET/CET-R, se pueden considerar en general satisfactorios. Sin embargo, se identifican varios aspectos que son susceptibles de mejora.
Palabras clave: *Tests, Uso de los tests, Evaluación de tests, Calidad de los tests, Impacto modelo evaluación de tests*

The Questionnaire for the Assessment of Tests (CET; Prieto & Muñiz, 2000) and the revised version of this questionnaire (CET-R; Hernández et al., 2016) have been applied systematically since 2010 by the Test Commission of the Spanish Psychological Association. The main goal is to provide practitioners with reliable information on the quality of the tests in order to improve test use. The aim of this paper is twofold. First, to present the results of the sixth review of psychological and educational tests, in which a total of 10 tests have been evaluated. Second, to assess the impact that the application of the CET/CET-R has had over these years on two key agents: test publishers and university lecturers who are responsible for training future test users. Both the results of the sixth review and the results of the survey to assess the impact of the CET/CET-R are satisfactory in general terms. However, some issues where there is room for improvement have been identified.
Key words: *Tests, Use of tests, Test Review, Test quality, Impact test review model.*

In the different areas of psychology, measurement and assessment using tests is a common practice. The results obtained from these measurements are used for different purposes such as to diagnose, make decisions, evaluate, advise or select; but, as indicated by the APA Standards (AERA, APA, and NCME, 2014), not all tests are well developed or used properly. Having information about the quality of a test and the appropriate use of its scores is crucial in the professional practice of psychology regardless of the disciplinary field of concern (clinical, educational, organizational, etc.). For several decades, different professional psychological organizations as well as associations and institutions have developed and promoted the use of guidelines or questionnaires to evaluate the quality of psychological and educational tests (Evers, 2012; Geisinger, 2012; Hernández, Ponsoda, Muñiz, Prieto, & Elosua, 2016), with the aim of improving the quality and rigor of tests. More specifically, since 2010, the Test Commission of the General Council of the Spanish Psychological Association has addressed the task of evaluating the tests published in Spain, in order to provide accurate and accessible information about quality of the tests in terms of their theoretical, practical and psychometric characteristics, so that it can be useful for practitioners in making informed decisions regarding the use of the tests (Hernández, et al, 2016). Prieto and Muñiz (2000) proposed a test review model that materialized through the CET Questionnaire (Test Evaluation Questionnaire) and in 2016 this instrument was reviewed by Hernández et al. (2016), resulting in the CET-R. During this time period, six reviews of the tests published in Spain have been carried out, including the one referring to the results presented in this article. In total 75 tests have been evaluated. The reports on these tests are available and can be downloaded from the Spanish Psychological Association (COP in Spanish) website in the corresponding entry to the National Tests Commission of the General Council of Psychology (https://www.cop.es/index.php?page=evaluacion-tests-editados-en-espana). A summary of the results of the first review can be found in Muñiz et al. (2011), the second in Ponsoda and Hontangas (2013), the third in Hernández, Tomás, Ferreres and Lloret (2015), the fourth in Elosua and Geisenger (2016) and the fifth in Fonseca-Pedrero and Muñiz (2017), all of which have been published in the journal *Papeles*

Received: 5 October 2018 - Aceptado: 9 November 2018
Correspondence: María Dolores Hidalgo. *Departamento de Psicología Básica y Metodología. Universidad de Murcia. Campus de Espinardo. 30100 Murcia. España. E-mail: mdhidalg@um.es*

del Psicólogo. In particular, the last revision (the one previous to this present one), describes the historical evolution of the review process of tests published in Spain (Fonseca-Pedrero & Muñiz, 2017).

All these reviews have been promoted by the General Council of the Spanish Psychological Association as an *information* strategy (Muñiz, Hernández, & Ponsoda, 2015), in order to provide test users with accurate and accessible information about the tests, and to respond to one of the main demands of psychology professionals: to have technical and psychometric information that helps them to make decisions based on evidence (Elosua, 2012; Muñiz & Fernández-Hermida, 2010). After five rounds of evaluations, six if we consider the one presented in this work; it is of interest to understand the impact that the model is having among the different fields and professionals in psychology. In this sense, although Fonseca-Pedrero and Muñiz (2017) provided data in terms of the number of reviewers and coordinators that participated in the review process, as well as the number of downloads of the reports available on the COP web page, we wanted to ask directly about the impact and use of these test quality reports.

Taking into account the above considerations, this article has two main objectives: 1) to present the results of the sixth review of tests published in Spain, and 2) to analyze the impact of the test review model (CET) and its CET-R review on academics (university professors who teach psychometrics and psychological assessment) and on the test publishers. The impact that the application of the test review model has had on applied professionals, beyond the aspects already considered by Fonseca-Pedrero and Muñiz (2017), remains to be known. This will be addressed in a larger study that will examine the opinion regarding tests in general (in line with the studies carried out by Muñiz & Fernández-Hermida, 2000, 2010).

Thus, first, after summarizing the review process followed, the results of the sixth test review are presented. Specifically, a total of ten tests have been reviewed: five for the measurement of mental and cognitive skills and intelligence, one for the assessment of depression, one for the assessment of problems in adolescents (anxiety, depression, self-esteem, psychosocial risk, etc.), one on solving mathematical problems, calculus and numeration, one on sensory processing and another for assessing attitudes and behaviors in medical patients. Specifically, the instruments reviewed have been the following: BADyG/E2-r, BADyG/S-r, CESPRO, BDI-FastScreen, MBMD, SENSORY PROFILE, Q-PAD, BAT-7, BPR and MATRICES (see Table 1 for a more detailed description). Secondly, the results are presented of the impact that both the evaluations carried out to date and the CET-R questionnaire itself are having among psychology academics. Finally, data are presented regarding the impact that this evaluation model is having on the test publishers.

SIXTH REVIEW OF TESTS PUBLISHED IN SPAIN

Description of the questionnaire

The CET-R questionnaire allows a qualitative and quantitative assessment of the quality of the tests, and its main objective is to provide test users with accurate and accessible information °about the quality of the tests reviewed.

The questionnaire is accessible and can be downloaded from the following web address http://www.cop.es/uploads/pdf/CET-R.pdf. In addition, along with the questionnaire itself, some general instructions are provided on how to answer the test and complete it, as well as a glossary of psychometric terms.

Although the main characteristics of the CET (Prieto & Muñiz, 2000) and CET-R (Hernández et al., 2016) can be consulted in the referenced works and in the different publications of the test

TABLE 1
LIST OF MEASURING INSTRUMENTS ANALYZED IN THE SIXTH TEST REVIEW

Acronym	Test	Publisher	Year of Publication
BADYG/E2-r	Batería de Aptitudes Diferenciales y Generales Renovado E2	CEPE, S.L.	2011
BADYG/S-r	Batería de Aptitudes Diferenciales y Generales	CEPE, S.L.	2011
BAT-7	Batería de Aptitudes de TEA	TEA ediciones	2015
BDI-FastScreen	Inventario de Depresión de Beck FastScreen para pacientes Médicos	PEARSON Education	2011
BPR	Batería de Pruebas de Razonamiento	TEA ediciones	2016
CESPRO	Batería para la Evaluación de las Estructuras Sintáctico-Semánticas que componen los enunciados de los problemas matemáticos y de la utilización de estrategias algorítmicas para su resolución	EOS	2016
MATRICES	MATRICES Test de Inteligencia General	TEA ediciones	2015
MBMD	Inventario Conductual de Millon para pacientes con diagnóstico médico	PEARSON Education	2014
Q-PAD	Cuestionario para la evaluación de problemas en adolescentes	TEA ediciones	2016
SENSORY PROFILE	Perfil Sensorial-2	PEARSON Education	2016

reviews carried out with this model, we will summarize the different aspects of psychometric quality that are considered. The CET-R includes three separate sections of test assessment: 1) General description of the test, 2) Assessment of the test characteristics and 3) Overall assessment of the test. The first section provides a brief description of the variable(s) measured, the area of application, the number of items and/or scales, the support for administration and correction, the required qualification for the use of the test in accordance with the documentation provided, the administration time, and even the price of the test. With regards to the second section, both the theoretical characteristics of the instrument (theoretical foundation, quality of materials and documentation, bibliography, and development of the items), and the more psychometric aspects such as the item analysis, validity and reliability evidence, and the scoring of the scale and interpretation of scores are assessed. The last section (overall evaluation of the test) requires a qualitative assessment of the instrument with special emphasis on its strengths and possible weaknesses.

Review process

The procedure followed in this sixth review is similar to that established in the previous ones. In the first place, the publishers selected the tests to be reviewed, and the Test Commission endorsed the proposal. On this occasion 10 tests were selected (one from the publishing house EOS, two from CEPE, three from Pearson and four from TEA). In addition, the Commission proposed the evaluation of different tests other than those proposed by publishers and used in schools. Thus, the evaluation of the test HABILMIND to assess learning, reasoning, reading, etc. abilities was proposed. Second, the coordinator appointed by the test commission (the first author of this article), selected a panel of experts for the review process, two per test. One expert had a more technical-psychometric profile and the other had a more theoretical profile, with knowledge and experience in the substantive aspects of the variable(s) measured by the test. The ethical aspects of the process were taken into account. As such, care was taken that the reviewers did not have a direct relationship with the authors of the tests, and that they had no conflicts of interest. In four cases, the reviewers initially selected refused to participate in the process for different justified reasons, so another reviewer had to be selected. Table 2 shows the final list of reviewers who agreed to collaborate as expert evaluators. We would like to highlight and show our appreciation for the work they carried out and their participation in the process.

The publishers made available to the General Council of the Spanish Psychological Association, free of charge, three complete copies of each test. From the General Council itself, one copy was sent to each of the two reviewers and the third was sent to the coordinator. The task of the experts, once the complete copy of the test was received, consisted of applying the CET-R to the assigned test, assessing each of the theoretical, practical and psychometric aspects. The reviewers received a symbolic amount of 50 euros (which some preferred to decline) in addition to a free copy of the test.

For each test a report was written which integrated the evaluations made by the corresponding reviewers. When there were discrepancies between the reviewers, the coordinator conducted an independent assessment based on the copies provided. Based on all this information, the quantitative assessments were assigned and the final assessment was made. Following the procedure established in previous evaluations, the report made on each test was sent to the publishers, so that both the editors and the authors of the tests could specify, clarify or comment on the aspects included in the evaluation report.

Once the comments and clarifications from the publishers were received, they were integrated into the final report, and the evaluations modified when this was justified. Finally, and prior to its publication on the COP website, the final report proposal was sent back to the publishers. Figure 1 shows a summary of the peer review procedure implemented.

Before proceeding to the results of the evaluation, it should be noted that the HABILMIND test could not be reviewed in the end since the test publishers did not make a free copy of the test available to the COP, despite several requests being made.

Results

The reports corresponding to each test that has been reviewed in this evaluation round, as we mentioned, can be consulted and downloaded at the following link

TABLE 2
REVIEWERS WHO HAVE PARTICIPATED IN THE SIXTH TEST REVIEW

Name	Affiliation
Maite Barrios Cerrejón	Universidad de Barcelona
Isabel Benítez Baena	Universidad de Loyola
Marcelino Cuesta Izquierdo	Universidad de Oviedo
Beatriz Delgado Domenech	Universidad de Alicante
José Pedro Espada Sánchez	Universidad Miguel Hernández
Eduardo García Cueto	Universidad de Oviedo
José Manuel García Fernández	Universidad de Alicante
Arantxa Gorostiaga Manterola	Universidad del País Vasco
Georgina Guilera Ferré	Universidad de Barcelona
Francisco Pablo Holgado Tello	UNED
Pedro M. Hontangas Beltrán	Universidad de Valencia
Urbano Lorenzo Seva	Universitat Rovira i Virgili
Luis Manuel Lozano Fernández	Universidad de Granada
Laura Nuño Gómez	Hospital Clínic de Barcelona
José Luis Padilla García	Universidad de Granada
Óscar Pino López	Hospital General de Granollers
Antonio J. Rojas Tejada	Universidad de Almería
Inés Tomás	Universidad de Valencia
María Soledad Torregrosa Díez	Universidad Católica San Antonio de Murcia
Ana Vanesa Valero García	Universidad de La Rioja
Carme Viladrich	Universidad Autónoma de Barcelona

https://www.cop.es/index.php?page=evaluacion-tests-editados-en-espana- This paper presents a summary of the main results obtained (see Table 3).

It can be seen that the ratings were good, since in most of the items of the questionnaire the tests were rated with scores of 3.5 or higher (rating from good to excellent). Of all of the tests and criteria evaluated, we found scores lower than 3 in only one test and for two criteria. The most valued and rated aspects of the tests refer to the criteria of *Materials and documentation* and *Theoretical foundation*. The next best rated criteria were the one relating to the reliability analysis of internal consistency and the one relating to the quality of the *Scales and interpretation of scores*. The tests evaluated provided different types of validity evidence (content, internal and relation with other variables), and were rated between good and excellent. Of the tests that were adapted versions (half of those reviewed), the average rating of the adaptation process was 4.2 (between good and excellent). In general, the adaptation process of these tests was carried out following the Guidelines of the International Test Commission (Hambleton, Merenda, & Spielberg, 2005, International Test Commission, 2018; Muñiz, Elosua, & Hambleton, 2013).

Compared to the previous reviews we found that, in this sixth review, 30% of the tests provided reliability evidence in terms of item response theory, providing information on the accuracy of the test in accordance with the subject's ability level. In addition, in 20% of the tests we found that validity evidence was provided in terms of studies of differential item functioning (DIF). These percentages are higher than in previous reviews, which suggests that these issues are being increasingly taken into account when publishing a test.

FIGURE 1
TEST REVIEW PROCEDURE

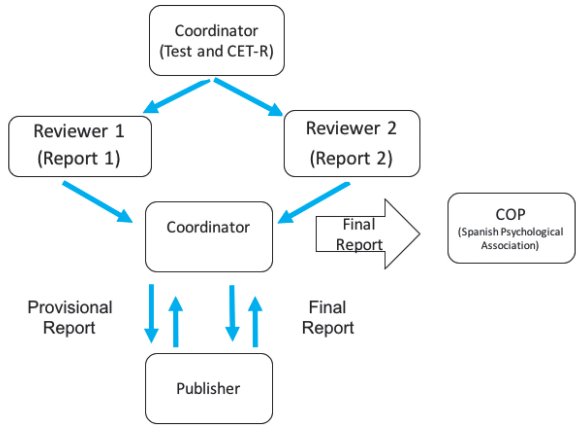

TABLE 3
SUMMARY OF THE SCORES OF THE TESTS ANALYZED IN THE SIXTH EVALUATION

Characteristics	BADyG/E2-r (2011)	BADYG/S-r (2011)	BAT-7 (2015)	BDI-FS (2011)	BPR (2016)	CESPRO (2016)	MATRICES (2015)	MBMD (2014)	Q-PAD (2016)	SENSORY PROFILE-2 (2016)
Materials and documentation	4.5	4.5	5	4.5	5	4	5	4.5	5	4
Theoretical foundation	5	5	5	5	5	4	5	5	3.5	3
Adaptation	-	-	-	4	5	-	-	4	5	3
Item analysis	4	4	5	4	4.5	4	5	-	3	-
Content validity	5	5	4	4.5	-	4	5	4	4.5	2.5
Validity based on relationship with other variables	3.2	3	4.1	4	4.3	4.2	4.6	3.7	4	2.7
Validity based oninternal structure	4	4	5	5	5	4	5	4	4.5	-
Validity based on DIF analysis	-	-	-	-	5	-	5	-	-	-
Reliability: equivalence	-	-	-	-	-	-	-	-	-	-
Reliability: internal consistency	5	4	5	4	5	5	5	3	5	3
Reliability: stability	-	-	-	4	3	-	5	3	3.5	3
Reliability: IRT	-	-	4.5	-	-	5	5	-	-	-
Reliability: inter-rater	-	-	-	-	-	-	-	-	-	4
Norms and interpretation of scores	4.7	4.7	4.3	4.5	4.5	4	5	4.5	4.7	3
Medium	Paper and Pencil Computerized	Paper and Pencil Computerized	Paper and Pencil Computerized	Paper and Pencil	Paper and Pencil	Paper and Pencil Computerized	Paper and Pencil*	Paper Pencil	Paper Pencil	Paper and Pencil Computerized (Online)
Correction automated by computer	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The scores of the table are given on a scale of five points whose values are: 1 = inadequate, 2 = adequate but with some deficiencies, 3 = adequate, 4 = good and 5 = excellent. When a hyphen appears (-) it means that no information is provided or it does not apply.
* Computerized adaptive version available

Finally, it is worth mentioning that most of the tests that have been reviewed have automated computerized correction (90%) and that more than half (60%) can be administered via computerized means.

IMPACT OF THE CET ON LECTURERS AND TEST PUBLISHERS

Survey of lecturers in psychometrics and psychological and (psycho)educational assessment

To determine the impact that the CET model and its revised version (CET-R) are having on the training of future psychology professionals in the fields of tests and assessment, the professors of psychometrics, psychological, and (psycho)educational assessment on psychology degrees at Spanish universities in the 2017-2018 academic year were surveyed. The main objective of this survey was to obtain information on whether the lecturers of psychology degrees that teach the subjects of psychometrics, psychological, and educational or psychoeducational assessment knew the CET model, whether they used it in their classes and, if so, in what way. It was also of interest to find out the reasons for not using this model as a teaching tool, when this was the case.

Procedure and Participants

Through the websites of the psychology degrees at Spanish

TABLE 4
IMPACT OF THE MODEL IN TEACHING BY TYPE OF SUBJECT - PSYCHOMETRIC OR PSYCHOLOGICAL ASSESSMENT OR (PSYCHO)EDUCATIONAL – AND OVERALL

	Psychometric (N=52)		Psychological and/or (psycho)educational ass.(N=41)		Overall[a] (N=95)	
Knowledge of the CET model before receiving the email requesting participation*	Yes 80.8%	No 19.2%	Yes 61.0%	No 39.0%	Yes 72.6%	No 27.4%
They did know the CET model	(N=42)		(N=25)		(N=69)	
They use the CET model for teachingConcrete use of the model	Yes 69.0%	No 31.0%	Yes 60.0%	No 40.0%	Yes 63.8%	No 36.2%
Specific use of the model[b]:						
Merely informative: so that students know that the model and its results can be useful when making decisions about the use of a test	44.8%		53.3%	47.7%		
Applied: the students have evaluated some test (s) applying the model	41.4%		20.0%	37.1%		
Applied: students have reviewed and/or worked in class with the reports available on the COP website))	10.3%		26.7%	15.9%		
Others: study rubric	3.4%		0.0%	2.3%		
Reasons for not using model[b]:						
Lack of time		61.5%		60.00%		56.0%
I do not think it contributes much to the training in the subject I teach		0.0%		10.0%		8.0%
It is not exhaustive enough		0.0%		0.00%		0.0%
It is too exhaustive		7.7%		10.0%		8.0%
Other: I had not thought about it but maybe I will use it in the future, I teach different evaluation techniques other than the tests, I used it once but it did not arouse interest among the students, it is not a competency for working on this subject.		30.8%		20.0%		28.0%
They did not know the CET model	(N=10)		(N=16)		(N=26)	
They would use the CET model for teaching[bc]	Yes or probably 70.0%	DK/NA 30.0%	Yes or probably 93.7%	DK/NA 6.3%	Yes or probably 84.6%	NS/NC 15.4%
Specific use that would be made of the model[b]:						
Merely informative: for students to know that the model and its results can be useful when making decisions about the use of a test	14.3%		13.3%		13.6%	
Applied: students would evaluate any test (s) applying the model	42.9%		13.3%		22.7%	
Applied: students would review/study in class with the reports available on the COP website)	14.3%		0.0%		4.5%	
I do not know yet	28.6%		73.3%		59.1%	

[a] Two lecturers were excluded from the comparison between types of subjects, since they taught both types (psychometrics and evaluation).
* Chi-square tests indicate that there are significant differences in the response patterns of teachers of both types of subjects.
[b] Due to the small sample sizes, no chi-square statistical contrasts were performed to compare response patterns according to the type of subject.
[c] The "no" option was not marked by any participant.
[d] DK/NA indicates "Don't know" or "Don't respond/Not applicable".

public and private universities, a total of 316 lecturers were identified who were invited to participate in the study by email. At the time of the survey, the total number of universities that taught the psychology degree was 56, with 53.57% public universities (30) and 46.43% private (26). Anonymity and confidential treatment of responses were guaranteed.

After several reminders were sent, a total of 97 lecturers (30.4%) responded to the survey. Of these, 73% belonged to public universities, half (50.5%) were male and the average number of years of teaching experience was 17.5 (SD = 11.1). After discarding the answers of two participants who indicated that they did not teach any of the subjects of interest (in spite of the information published on the websites of their universities), the majority (49.5%) were lecturers in psychometrics, followed by 35.8% who were lecturers in psychological assessment. Only two participants (2.1%) indicated that they were lecturers in educational or psychoeducational assessment. The rest (12.6%) indicated that they taught several of the aforementioned subjects or combined some of these subjects with others outside of assessment itself. The responses of the 95 professors were used to give the overall results. In order to obtain differentiated information according to the main teaching subject, two groups were formed: those who had psychometrics as their only or main subject (if they taught a second subject, this had nothing to do with assessment topics - for example, they taught statistics) and those who had as the only or main subject one or several subjects related to assessment -psychological and/or (psycho) educational. Those who taught both psychometrics and assessment subjects were not included in the differential analyses.

Results

The results are presented in Table 4, both globally and differentiated according to the type of subject taught (psychometrics vs. psychological and/or (psycho)educational assessment). Specifically, most of the professors who responded to the survey (72.6%) said they knew the CET model, while 27.4% said they did not know it. The percentage of people who knew the model is significantly higher for professors of psychometrics compared to those of evaluation (χ^2=4.46; g.l.=1; p<.05).

Of the total number of lecturers who said they knew the model, the majority (63.8%) do use it in their classes at some time. In this case, no statistically significant differences exist depending on the type of subject taught. In general, the use that is given is mostly informative (in 47.7% of cases), so that students know that the model and its results can be useful when making decisions about the use of a test. There is, however, also a more applied use: either the students evaluate any test(s) applying the model (37.1%) or they review or work with the reports available on the COP website (15.9%). In this case, no statistical test was carried out, due to the small sample sizes for relying on the chi-square results. However it is observed that the largest difference in the way the model is used in psychometric or assessment classes is found in the evaluation of tests applying the CET model, which turns out to be a more frequent practice in the subject of assessment than in that of psychometrics.

Among those lecturers who do not use the CET model for teaching, despite knowing it, the majority indicates that they do not do so due to lack of time (56%), this being the reason most offered by both professors of psychometrics and psychological and (psyco)educational assessment.

Of the total number of lecturers who did not know the model until they were invited to participate in the study, the majority note that, in the future, they will use it for teaching (42.3%) or they probably will (42.3%), while 15.4% still do not know. In addition, the majority (59.1%), especially among teachers of psychological and/or (psycho) educational assessment (73.3%), were not clear about what specific use they would give the model when teaching the subject, considering it necessary to consult in more detail the model and the evaluations carried out.

TABLE 5
LIST OF QUESTIONS THE PUBLISHERS WERE ASKED

Questions asked
1. Indicate what the impact of the test quality review model has been on the processes of development and adaptation of the psychological and educational tests published by your publisher. Highlight the most positive aspects in the creation and adaptation of new tests in your publishing house.
2. Highlights which aspects of the model and its application could be modified to improve the quality of the tests in the development and publication process.
3. Indicate what measures could be taken to increase the impact of the model and for the positive evaluation of a test by the commission of tests to be interpreted by users as a seal of quality.
4. If you think that the model is having a counterproductive effect on publishers and users, briefly describe what these effects are.

Survey of the main Spanish test publishers

The quality of the tests depends directly on the publishers that publish them and their commitment to maintaining high quality standards. The involvement and good disposition of the main test publishers in Spain (TEA Ediciones, Pearson, EOS and CEPE) has been essential in implementing the application process of the CET. It should also be noted that publishers play a crucial role in ensuring that tests have adequate quality criteria and are only acquired by qualified professionals. To find out the opinion of these publishers about the impact of the CET model on the quality of the tests they publish, to understand how they would improve the application of the model and know what measures they would take for the evaluations to have a greater impact, representatives of the four publishers mentioned were surveyed, with a response being obtained from three of them. Specifically, we asked them four questions (see Table 5), and the most notable aspects of the answers obtained are summarized below.

In general, when asking about the impact of the CET model on the processes of test development and adaptation, the comments were highly positive. They emphasize that the CET model is taken as input by authors and publishers in the development and adaptation of the tests, as well as for the publishing of the test manual. In fact, in the manuals, the evidence that supports the technical quality of the tests has been presented in a much more detailed way in recent years. For example, some publishers, in addition to the reliability values of different samples and scales, include the averages obtained across samples, which is more summarized information, as well as the mark obtained, according to the quality criteria established by the CET model (good, excellent, adequate, etc.). They also value very positively that the model has become a system of continuous improvement that reinforces the quality standards that publishers seek. Also, for the potential user, it means greater confidence when selecting and using a particular test compared with others. An additional interesting aspect related to the other major agent analyzed in the present study, the university lecturers, is that, according to some of the participating publishers, the universities have moved from buying more traditional tests, which are in some cases obsolete, to buying tests that are less well-known, but widely used in the applied context, updated, and supported by empirical evidence on technical and psychometric quality.

Several issues are mentioned with regards to the aspects of the model and its application that could be modified to improve the process of development and publication of the tests. Since the manuals are made much more detailed and exhaustive, this makes the result more expensive and, in some cases, adds complexity for test users. One possibility is to stick to the most basic information in the manual, and publish the additional detailed evidence in the form of complementary online materials. These materials would also be considered by the commission when evaluating the quality of the tests, the results of which are published on the COP's website. Similarly, it is also indicated that different types of evidence on the quality of the test could be prioritized according to the purpose of the test and it could be determined which aspects are essential in order to publish a test and, if appropriate, which aspects are not. Finally, the need is noted to include in the review process other tests that are not published by the best-known publishers, which are those that have representation in the test commission. Otherwise, the tests of these latter publishers could be affected negatively compared to other tests that are not evaluated, but are widely used in different areas, without rigorous empirical studies to support their quality. Only the tests that are evaluated are exposed to criticism, so the review should be opened to any test that is being used professionally. This would help improve the quality of the tests that are developed and adapted, regardless of the publisher behind them. According to some publishers, it is a contradiction that publishers with a certain amount of experience in evaluating and technically improving the tests, who are giving the green or red light to the works that are published, are now receiving evaluations, in some cases negative ones, by professionals far from the more applied fields of psychometrics.

As regards the measures that could be taken by the commission to increase the impact of the CET and the evaluations carried out, several aspects are noted. In addition to the suggestion to include in the review processes other tests that do not depend on the publishers with representation in the test commission, a minimum score or a percentage of criteria in which excellence is achieved could be established, serving as a cut-off point to receive a seal of quality from the commission. The tests that do not reach sufficient quality guarantees for certain purposes should also be clearly indicated. Finally, in order for the evaluations to have a real impact on the professionals, there is a need to reduce the distance between the academics (who are mostly involved in the evaluation processes of the tests under review, but who sometimes have never participated in test development or adaptation processes) and the applied professionals (the final recipients of the reviews). As it is indicated, the distance between the academic and the applied sides means that the process is seen as somewhat distant to the reality of the practitioners, like something too technical that has little to do with the real needs and the response that the tests analyzed provide to those needs. To reduce this gap, the continuous training and updating of the practitioners is necessary, for them to be able to understand the aspects of the evaluation that are more related to the psychometric and technical advances. However, it could also be useful to increase the number of applied professionals with contrasted psychometric knowledge, who participate in the review process.

These last considerations are related to some of the possible counterproductive effects of the model and its application. Within the possible negative effects, it is mentioned that "more" is not always synonymous with "better". For example, in the case of norms, the model gives a higher score when a wide range of norms is used instead of a single norm which is suitable for the target population. According to some publishers this could lead to additional norms being introduced artificially that may be unnecessary or not useful, in order to obtain a better score in the CET. In addition, it is also possible to embrace the practice of giving information on all the aspects that are considered in the model, even when they are not relevant for the use that will be made of the test. It is also noted that some of the criteria are excessively demanding in realistic applications (for example, the sample sizes required for test-retest reliability, especially in certain types of populations). Finally, the subjectivity inherent to the valuations that are given, especially the qualitative ones, is also valued as a negative aspect in the process. To qualify this subjectivity to a certain extent, the role of the coordinator that integrates the evaluations of the different reviewers of each test is crucial, but the process should be more demanding and urge test reviewers to provide a sound justification for their assessments.

All these issues should be progressively taken into account in future evaluations in order to continue improving the CET, its impact on the improvement of the tests and on their use by the professionals and academics who teach subjects related to tests and assessment through tests.

CONCLUSIONS

In this sixth test review, a total of ten instruments have been evaluated with the CET-R. Both the quantitative and qualitative evaluations were good, the tests being evaluated with a score of good to excellent for most of the items of the CET-R. The tests reviewed, in general, provide reliability and validity evidence of the use from the scores obtained. In addition, in this test review it was found that almost one third of the instruments analyzed had addressed the accuracy of the measurement based on IRT models. Item analysis using IRT provides valuable information that helps the development and improvement of the instrument under construction. The item characteristic curves (ICCs), as well as their corresponding item information functions (IIFs), are graphic tools that show the level of ability/trait where the items are more precise. These toolsguide us in the selection of the items that allow us to reach the minimum quality necessary in terms of reliability (Ames & Penfield, 2015; Hidalgo & French, 2016). This is especially useful in making decisions, when we have to be precise at a certain level of ability that has been established as a criterion or cut-off point for the interpretation of the score. The use of these measurement models, both theoretically and practically, allows us to obtain useful arguments regarding the valid use of the scores. Added to what has already been mentioned, we can see important advantages in the use of these models, as the basis for the construction of adaptive tests (Olea, Abad, & Barrada, 2010) and a more ecological assessment. Although these are models that generally require larger samples (Muñiz, 2010; Penfield, 2014), the advantages mentioned mean that the increase in the tests that use IRT in the current test evaluation, in comparison with the previous rounds of test evaluations, is viewed as very positive.

In this sixth review of tests published in Spain, it has also been shown that more than half of the tests evaluated (60%) can be administered via computerized means. Although it seems that, in the clinical and educational fields, paper and pencil is preferred over computerized administration (Antón, 2017), it would be interesting to ask psychology practitioners about the reasons for these preferences, taking into account the possibilities of new formats of items that the advances in information and communication technologies provide us. We also wish to emphasize the importance of sharing databases of reliability and validity studies of tests, and the need to create a repository with this type of data in order to accumulate evidence of validity and reliability. As Botella and Ortego (2010) indicate, sharing direct data from research would result in a more efficient accumulation of knowledge, which in the case of the metric quality of the tests would contribute to their improvement.

Furthermore, in any context (organizational, educational, health, etc.) where a quality assurance system has been implemented, the evaluation of the quality system itself is key in the improvement cycle. In this sense we will comment on some aspects related to the implementation of the CET-R that have been highlighted in the process of test evaluation of this sixth review. In the first place, the need to facilitate the application of the CET-R, both for the reviewers and for the coordinator him- or herself, has been confirmed when integrating the results of the evaluations received. Thus, computerizing the CET-R (computerized application) would speed up its application and allow the incorporation of more technical-psychometric information, as well as concrete examples of use and/or video tutorials, accessible online, to reiterate some of the suggestions of Fonseca-Pedrero and Muñiz (2017). The computerized version should also facilitate the obtaining of the numerical summary-assessment for each final characteristic contemplated in the CET-R. At the same time, it could be more useful as a self-assessment tool for those researchers or professionals immersed in the process of developing a test.

Secondly, some difficulties have been detected when evaluating questions related to sample size. An example of this would be in reliability estimates such as internal consistency or test-retest. Although the different response options for these items are clear, there are problems in integrating the evaluations of those instruments in which there are different versions of the questionnaire (levels) depending on the age of the people assessed, and for which the reliability evidence has been obtained in different sample sizes for each level, but which do not reach a large sample (N>500). The possibility of several studies with a moderate sample and other small ones is not explicitly included in the CET-R. On the other hand, it is especially difficult to assess the sample size in the test-retest reliability studies of a test when it has different versions (levels) and different sample sizes are used in each version. The application of the CET-R could be improved by including examples of application that guide the reviewer in the evaluation process. Thirdly, it is recommended that in the 2.11.2.2.6 criterion the response graduation (Good, Very good and Excellent) is modified in the most extreme options, as in this case the graduation would stay the same as in other questions of the questionnaire (Good and Very Good). Fourth, for those evaluated tests in which the evidence of reliability and/or relationships with other variables have been obtained in several samples of different sizes, it would be of interest to consider using an average estimator of the reliability coefficient or the correlation coefficient weighted by the sample size. We know that some of these issues are already being taken into account in the seventh test review.

Looking back, the usefulness of the CET-R as a tool for evaluating the quality of the tests has been made clear through the already 75 tests evaluated. In addition, it has been suggested that this model is having an impact positive both in the academic sphere related to the teaching of psychology, and

in that of the test publishers (e.g. Hernández et al., 2015, 2016). In this study, this impact has been analyzed in a more systematic way. On the one hand, the survey responses of university lecturers regarding the use of the CET (and its revised version CET-R) in the teaching of psychometrics, psychological and educational assessment and related disciplines have been analyzed. On the other, we have interviewed the representatives of the test publishers in the test commission, to understand the impact of the model in the construction/adaptation of tests, and to understand which aspects of the model implementation could be improved.

Regarding the impact of CET/CET-R in the field of university teaching in psychology, although the percentage of lecturers who responded to the survey only represents 30.4% of the total that were invited to participate, the results show promising data. First, two thirds of the teachers who responded to the survey knew the evaluation model. Although this figure is positive, dissemination work can still be improved to reach 100% of the teaching staff. In this sense, the action taken in contacting the lecturers to ask them about the CET/CET-R may contribute to increasing this percentage in the future. It is also a good result that the majority of professors who knew the CET/CET-R did use it in their classes at some point (63.8%). This is a way for future professionals to learn about the model and the results of the evaluations. Among those who do not use the model despite knowing it, lack of time is the reason most frequently offered. Finally, we also take as positive data the fact that, after the survey, of the total number of lecturers who did not know the model until they were invited to participate in the study, approximately 85% report that they will use it in their classes, either definitely or probably.

With regards to the impact that the CET/CET-R model has had on the test publishers, there are several aspects to be highlighted: (1) The CET-R model is taken as a guide by both authors and publishers in the process of development and adaptation of the test, as well as in the editing and preparing of the test manual, and (2) the model is established as a benchmark system for continuous improvement to achieve the quality standards of the tests. However, (3) it is necessary to include in the review process other tests that are not published by the best-known publishers and that are used in different professional fields. Finally, (4) so that the test reviews have a real impact on practitioners, it is necessary to reduce the distance between the academic (the technical) and the applied spheres. Thus, establishing effective strategies to disseminate the evaluation reports of the quality of the tests among applied professionals, is a challenge to be addressed by the Test Commission. Combining the technical-psychometric rigor of the test evaluation reports with activities that make this knowledge accessible to practitioners is a task yet to be undertaken. Moreover, the continuous training and updating of practitioners is also necessary in order for them to understand the evaluation aspects that are most related to psychometric and technical advances. The dissemination of the CET-R in professional areas of clinical and educational intervention (mental health centers, hospitals, educational centers, etc.) will surely be a contribution in this direction and will result in a better use of the tests.

CONFLICT OF INTERESTS

There is not conflict of interests

REFERENCES

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.

Ames, A.J., & Penfield, R.D. (2015). A NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice, 34,* 39-48.

Antón, M. (2017). Mesa redonda "La salud de los test en España" [Round table "The health of tests in Spain"]. Coordinator: Ana Hernández. *III Congreso Nacional de Psicología*. Oviedo, 3-7 July.

Botella, J., & Ortego, C. (2010). Compartir datos: hacia una investigación más sostenible [Sharing data: towards more sustainable research]. *Psicothema, 22,* 263-269.

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.

Evers, A. (2012). The Internationalization of Test Reviewing: Trends, Differences, and Results. *International Journal of Testing, 12,* 136-156.

Fonseca-Pedrero, E., & Muñiz, J. (2017). Quinta evaluación de tests editados en España: Mirando hacia atrás, construyendo el futuro [Fifth evaluation of tests published in Spain: Looking back, building the future]. *Papeles del Psicólogo, 38*, 161-168.

Geisinger, K. (2012). Worldwide test reviewing at the beginning of the twenty-first century. *International Journal of Testing, 12,* 103-107.

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

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G., & Elosua, P. (2016). Revisión del modelo para evaluar la calidad de los tests utilizados en España [Assessing the quality of tests in spain: Revision of the spanish test review model]. *Papeles del Psicólogo, 37,* 192-197.

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.

Hidalgo, M.D., & French, B. (2016). Una introducción didáctica a la Teoría de Respuesta al Ítem para comprender la construcción de escalas [A didactic introduction to item response theory for understanding the construction of scales]. *Revista de Psicología Clínica con Niños y Adolescentes, 3*, 1-9.

International Test Commission (2018). ITC Guidelines for Translating and Adapting Tests (Second Edition). *International Journal of Testing, 18*, 101-134.

Muñiz, J. (2010). Las teorías de los tests: Teoría Clásica y Teoría de Respuesta a los Ítems [Test theories: Classical Theory and Item Response Theory]. *Papeles del Psicólogo, 31, 57-66*.

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

Muñiz, J., & Fernández-Hermida, J.R. (2000). La utilización de los tests en España [The use of tests 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,* 108-121.

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

Muñiz, J., Hernández, A., & Ponsoda, V. (2015). Nuevas directrices sobre el uso de los tests: investigación, control de calidad y seguridad [New guidelines for test use: research, quality control and security]. *Papeles del Psicólogo, 36*(3), 161-173.

Olea, J., Abad, F.J., & Barrada, J.R. (2010). Tests informatizados y otros tipos de tests [Computerized tests and other new types of test]. *Papeles del Psicólogo, 31,* 97-107.

Penfield, R.D. (2014). An NCME Instructional Module on Polytomous Item Response Theory Models. *Educational Measurement: Issues and Practice, 33,* 36-48.

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 to evaluate the quality of tests used in Spain]. *Papeles del Psicólogo, 77,* 65-72.