Papeles del Psicólogo / Psychologist Papers, 2021 Vol. 42(1), pp. 1-9
https://doi.org/10.23923/pap.psicol2020.2937
http://www.papelesdelpsicologo.es
http://www.psychologistpapers.com

EIGHTH REVIEW OF TESTS PUBLISHED IN SPAIN: A PARTICIPATIVE EXPERIENCE

Carme Viladrich[1], Eduardo Doval[1], Eva Penelo[1], Joan Aliaga[1], Albert Espelt[1,2], Rebeca García-Rueda[1] and Ariadna Angulo-Brunet[1]
[1]*Universitat Autònoma de Barcelona.* [2]*CIBER de Epidemiología y Salud Pública*

La Comisión de Test del Consejo General de la Psicología en España promueve anualmente la revisión de la calidad de diferentes test publicados. Este trabajo tiene un doble objetivo: a) presentar los resultados de la octava edición y b) considerar la aportación de la universidad en dicho proceso. En esta edición participaron 10 especialistas, 332 estudiantes y siete profesores, adaptándose el protocolo estándar de revisión al formato aprendizaje-servicio. En cuanto a los resultados, la calidad de los 11 test evaluados fue adecuada (promedio de 3,9 puntos en una escala 1-5) y similar a años anteriores (r = 0,90). El desarrollo y la baremación fueron puntos fuertes, mientras que se proponen mejoras en otros aspectos. El aprendizaje-servicio contribuyó a la diversificación de voces en el proceso observándose una calidad similar entre los informes del estudiantado y los emitidos por especialistas y un grado de acuerdo esperable (r = 0,67) entre ellos. Concluimos que el presente proyecto ha permitido identificar la oportunidad de profundizar en el uso de lenguaje compartido para fortalecer la comunicación entre las casas editoriales, la comisión promotora del modelo español de revisión de test, y las personas usuarias de los test, particularmente si se trata de principiantes.
Palabras clave: *Evaluación de test, Calidad de los test, Psicometría, Aprendizaje-servicio*

Every year, the Test Commission of the Spanish Psychological Association promotes the assessment of the test quality of several published tests. The aim of the present study is two-fold: a) to present results for the eighth review, and b) to consider the contribution of the universities in this process. Ten experts, 332 students, and seven professors participated in this edition and the standard protocol for review was aligned towards a service-learning format. For the 11 tests assessed, results showed an adequate quality (average of 3.9 points on a 1-5 rating scale) similar to previous years (r = .90). The strengths were test development and standardization, and there were a number of proposals for improving other sections suggested. The service-learning approach contributed to the diversification of voices in the process, with students' and experts' reports showing similar quality and an expected level of agreement (r = .67). We conclude that this project has helped to identify the opportunity to further expand the use of shared language in order to strengthen the communication between the test publishers, the promoters of the Spanish model of test assessment, and the test users, especially in the case of beginners.
Key words: *Test review, Test quality, Psychometrics, Service-learning.*

The test review promoted by the Test Commission of the Spanish Psychological Association (COP, http://cop.es) in collaboration with the Test Commission of the European Federation of Psychologists' Associations (EFPA, http://efpa.eu; Evers et al., 2013) began in 2011 (Muñiz et al., 2011) and since then 84 reviews have been published at the approximate rate of one edition per year. This initiative is intended to respond to the needs for independent information by the profession (e.g., Hidalgo & Hernández, 2019) and for training (e.g., Fonseca-Pedrero & Muñiz, 2017), which are more relevant when the tests are used for making decisions with important consequences for the individuals being assessed (e.g., Hernández et al., 2015).

Received: 27 abril 2020 - Accepted: 8 junio 2020
Correspondence: Eva Penelo. *Facultat de Psicologia. Universitat Autònoma de Barcelona. C. de la Fortuna s/n. 08193 Bellaterra (Cerdanyola del Vallès). España.*
E-mail: eva.penelo@uab.cat

Among the actors involved in the process, the Test Commission of the COP, made up of professionals from the academic world and the test publishers, has assumed the function of prioritizing the tests to be evaluated in each edition. In turn, the profile of the majority of participants in the reviews has been one of people who occupy senior positions in different academic specialties in the fields of psychology and education. Once the process was consolidated, the willingness to include new voices in the review was made explicit (Elosua & Geisinger, 2016), particularly those who speak from the professional field and from junior positions (Fonseca-Pedrero & Muñiz, 2017). A further step in this direction would be the extension to students who will join the profession shortly, under the tutorship of their teachers. In addition to integrating a new voice into the process, this strategy would give professors the opportunity to combine review and training tasks, at least in some academic specialties.

The potential of the European model of test reviewing as a

training tool has been widely recognized by university professors (Hidalgo & Hernandez, 2019; Vermeulen, 2019). This was also seen in the subject of Psychometrics at the *Universitat Autònoma de Barcelona* (UAB, http://uab.cat), in which from the academic year 2011/12 onwards we implemented a problem-based learning project using the test evaluation model promoted by the COP. The student writes an evaluation report of a psychological, educational, or speech-language pathology test, by completing the Revised Test Evaluation Questionnaire (CET-R; Hernandez, et al., 2016) and defends it orally as part of the evidence of learning presented (Doval et al., 2013; Viladrich, Doval, Aliaga et al., 2014). Over the last decade, our students have evaluated a total of 91 tests chosen by their teachers from among those available in our collection, we have presented favorable data on the validity of their reviews in comparison with those of specialists (Viladrich, Doval, & Penelo, 2014), we have studied the effect of early adherence to the project on academic results (Espelt et al., 2016), and we have contributed to the revision of the Spanish model (Hernández et al., 2016), in addition to collaborating individually as reviewers in different editions. During the 2019/20 academic year, we accepted the challenge of leading the eighth edition of the review of tests published in Spain. To do this, we have adapted the teaching methodology we had been using to a service-learning format (ApS, Redondo-Corcobado & Fuentes, 2018) that has gone beyond the university to address the entire professional community (Viladrich et al., 2019, 2021).

In this edition, the COP Test Commission commissioned us to review 11 tests aimed at measuring intelligence, verbal skills, and personality, published between 2006 and 2019. Specifically, these are the six levels of the BADyG test, and the BRIEF-P, CELF-5, MCMI-IV, PECO, and TONI-4 tests. More details of all of them can be seen in Table 1. Consequently, the first objective of this article is to present and discuss the quality evidence of the 11 tests submitted for evaluation in the eighth edition of the review of tests published in Spain. The second objective is to present and discuss the contribution of the university in relation to two new roles: as a proponent of the tests to be evaluated and as a participant in the review process through the psychology student body under the guidance of the teaching staff.

METHOD

Participants

There were 332 student participants (78.9% female) who formed 69 work teams and the teaching staff of the Psychometrics course, which is compulsory for the third year

TABLE 1
TESTS EVALUATED IN THE EIGHTH EDITION

Acronym	Name	Publisher	Year of publication
BADyG/i	Batería de Actividades mentales Diferenciales y Generales, Nivel infantil CEPE, S.L. [Differential and General Mental Activity Battery, Child Level]	2019	
BADyG/E1-r	Batería de Actividades mentales Diferenciales y Generales, Nivel E1 renovado [Differential and General Mental Activity Battery, Level E1 renewed]	CEPE, S.L.	2019
BADyG/E2-r	Batería de Actividades mentales Diferenciales y Generales, Nivel E2 renovado [Differential and General Mental Activity Battery, Level E2 renewed]	CEPE, S.L.	2019
BADyG/E3-r	Batería de Actividades mentales Diferenciales y Generales, Nivel E3 renovado [Differential and General Mental Activity Battery, Level E3 renewed]	CEPE, S.L.	2019
BADyG/M-r	Batería de Actividades mentales Diferenciales y Generales, Nivel M renovado [Differential and General Mental Activity Battery, M-Level renewed]	CEPE, S.L.	2019
BADyG/S-r	Batería de Actividades mentales Diferenciales y Generales, Nivel S renovado [Differential and General Mental Activity Battery, S-Level renewed]	CEPE, S.L.	2019
BRIEF-P	Evaluación Conductual de la Función Ejecutiva- Versión infantil [Behavioral Assessment of Executive Function - Children's Version]	TEA Ediciones	2016
CELF-5	Evaluación Clínica de los Fundamentos del Lenguaje, 5 [Clinical Evaluation of Language Fundamentals, 5]	Pearson Educación	2018
MCMI-IV	Inventario Clínico Multiaxial de Millon, IV [Millon, Clinical Multiaxial Inventory IV]	Pearson Educación	2018
PECO	Prueba para la Evaluación del Conocimiento Fonológico [Phonological Knowledge Assessment Test]	EOS	2006
TONI-4	Test de Inteligencia No Verbal- 4 [Non-Verbal Intelligence Test - 4]	TEA Ediciones	2019

of the degree in Psychology at the UAB. Furthermore, six reviewers and four specialists in Psychometrics, health, or education participated from different Spanish institutions (see upper left of Table 2), as well as an unspecified number of people from each test publisher.

Instruments

CET-R. The quality criteria of the Spanish model of test evaluation are reflected in the CET-R (Hernández et al., 2016), which consists of three sections. In the first, the characteristics of the test are described; in the second, its properties are evaluated; and in the third, all the evaluations are summarized. The properties of the test are evaluated by answering closed questions with five response categories ordered from insufficient to excellent (10 questions about the development of the test, 18 about validity, 14 about reliability, and nine about the interpretation of the scores). In addition to being based on a subject heading, these assessments are discussed in several open-ended questions. Since not all tests require the same quality evidence, the model is made more flexible by subjecting the applicability of each type of evidence—particularly in the sections on reliability and interpretation of scores—to the judgment of whoever evaluates a particular test.

Student performance and satisfaction. These are provided as standard for all university subjects. Retention and

TABLE 2
PARTICIPANTS IN THE EIGHTH TEST REVIEW

Professional reviewer (affiliation)	Editor-Tutor (Universitat Autònoma de Barcelona)
Alejandro Veas Iniesta (Universidad de Alicante)	Albert Espelt
Ana Isabel González Contreras (Universidad de Extremadura)	Ariadna Angulo-Brunet
Gerardo Aguado Alonso (Universidad de Navarra)	Carme Viladrich
Maria Dolores Gil Lario (Universitat de València)	Eduardo Doval
Maria Dolores Prieto (Universidad de Murcia)	Eva Penelo
Marijo Garmendia Txapartegi (Zarauzko Berritzegunea)	Joan Aliaga
Miguel Angel Carbonero (Universidad de Valladolid)	Rebeca García-Rueda
Montse Bartroli Checa (Agència de Salut Pública de Barcelona)	
Natalia Hidalgo-Ruzzante (Universidad de Granada)	
Rafael Martínez Cervantes (Universidad de Sevilla)	
Student reviewer (Universitat Autònoma de Barcelona)	
Ainoa Barreiro Escobar	Laura Saavedra García
Alba Rodríguez-Delgado	Layla Ishak-Tello
Aleix Jané-Alsina	Mar Viniegra Pintado
Anna Orts	Maria Peiro
Cristina Fuste-i-Valentí	Maria Silva Pereira
Carmen María Segura Sanchez	Marian Granados-Gamito
Daniel Steinherr Zazo	Marina Clivillé Domingo
Dunia Hanafi-Alcolea	Marta Valera-Guiot
Fátima Zarfani Azarfane	Meritxell Bagué Solé
Fernando Mengual-Rodas	Meritxell Barroso Cantero
Gemma Casimiro Fernandes	Mireia Gamez-Broto
Gemma Lapeira-Casé	Natalia Llobet-Vallribera
Isaac Pardo-Niño	Núria Coma Bravo
Joan Martínez-Vidal	Oriol Martín-Corella
Judit Reyes Griñena	Paula Jimenez-Ventura
Judith Moya	Queralt Mas-Jarque
Júlia Bartra Pallarès	Raquel Villar Mateo
Júlia Carrasco Hernández	Roser Bigorra Fargas
Ksenia Ouzioumova	Sílvia Solano Selvas
Laia Cervigón Moreno	Xenia Pla-Ruiz

success rates are assessed (UAB, n.d.-a), as well as the perceived strengths and weaknesses of the subject (UAB, n.d.-b).

Procedure

Figure 1 shows the procedure of obtaining two independent reviews, which are reconciled by the editor who then considers the comments of the publishers before writing and disseminating the final report and the evaluation process. See Gómez (2019) for more details on standard protocol. For our part, we have developed a specific protocol to apply to the ApS academic project. Each of the seven authors of this article acted as editor of one or two of the tests to be evaluated until the 11 that formed the assignment were covered. The first author also channeled all the information flows with the Test Commission, reviewers, and publishing houses and supervised all the reports. During the academic year, in each Psychometrics practice group, one of the tests was evaluated in the form of a competition among four or five teams made up of three to six people. Each team developed a draft, received comments from their tutor, and then wrote the final report. The winning report of each test was considered as Review 2 (see authors in the bottom part of Table 2). In the event that the award was considered void, the tutors assumed this role. Subsequently, the professors conducted two discussion sessions to establish a common editorial line and assess the project.

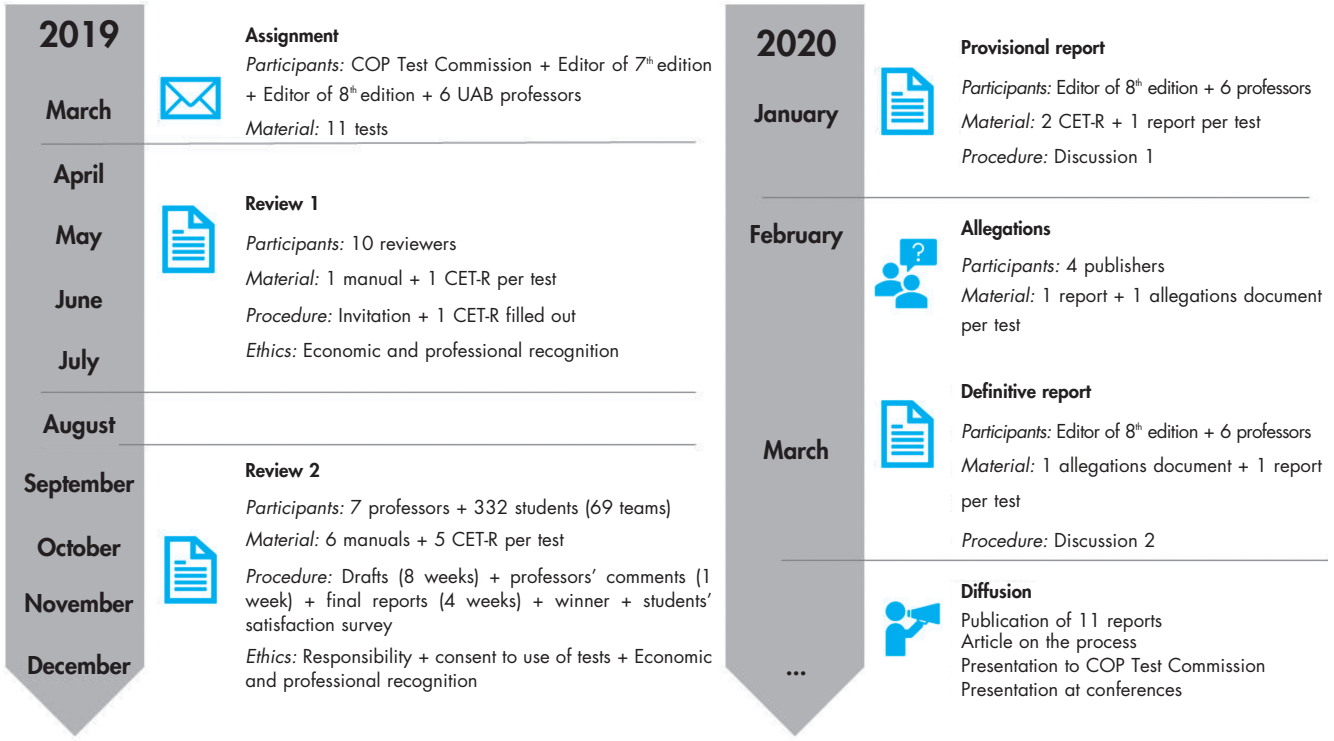
In terms of ethical precautions, the adaptation to the ApS format was as follows. Each member of the teaching team agreed to participate in the project before the start of the course; alternatively, they could restrict their activity to the usual tutoring of their students. For the students it was a mandatory and evaluable activity as it had been for the last decade. They were informed about the project during the first class session as well as permanently in the virtual classroom. Each student acknowledged by their signature that they had read the document on copyright laws affecting the materials. Furthermore, each signatory of a winning project gave their written consent to use their text as Revision 2 and to publish their name in the dissemination process. The financial reward offered by the COP to the editor and Reviewer 2 was paid into a UAB fund to cover the costs of the Psychometrics course.

RESULTS AND DISCUSSION

Quality of the Tests Evaluated

The review reports of the 11 tests evaluated are the main result of this work and can be consulted and downloaded from the COP website, under the section corresponding to the year 2019 (https://www.cop.es/index.php?page=evaluacion-tests-editados-en-espana). As a summary, Table 3 shows the scores of each of the 11 tests evaluated in the aspects related to the development of the test, the validity evidence, the

FIGURE 1
PROCEDURE OF TEST REVIEW IN THE EIGHTH EDITION



estimation of the reliability of the scores, and the scoring and interpretation of the scores.

First, we emphasize that the tests evaluated in this edition were developed and scaled very appropriately, as can be seen from the first and last categories in Table 3. In all the manuals there was a theoretical foundation between good and excellent, and the same was true of the adaptation process. Although with a little more variability, the materials and documentation of the tests were also evaluated in this same range, as well as the large majority of the studies of norms. These results are similar to or higher than those obtained in the other reviews based on the CET-R (Gómez, 2019; Fonseca-Pedrero & Muñiz, 2017; Hidalgo & Hernández, 2019) and, as a whole, they illustrate the robustness of the tests that publishers submit to evaluation year after year in these aspects. It should be noted that the presentation of data from the item analysis did not always reach acceptable values according to the CET-R heading. In this sense, we recover here the suggestion of Ponsoda and Hontangas (2013) that in the manuals additional materials could be mentioned that are available on the website of the publisher, as is already done, for example, with the BADyG test battery. With regard to the interpretation of scores, the most significant improvement is that all manuals justified the published reference points as an aid to decision-making. This is the recommendation of the CET-R (section 2.13.2.1), and an example of good practice is the sensitivity and specificity data published in the MCMI-IV test manual. At the very least, an effort should be made in all manuals to clarify that the fact that if a person occupies an atypical position in relation to his or her normative group, this does not in itself have clinical significance. This would help correct possible misuses that can be reasonably anticipated (American Educational Research Association [AERA] et al., 2014, standard 7.1).

The information contained in the section on validity was qualified as adequate, which means that the use of the tests evaluated gained the confidence of those who reviewed them, although it is undoubtedly the section with the greatest margin for improvement. A first opportunity for improvement would be to associate with each of the proposed uses of the test the validity evidence that supports it. This was recommended in the review signed by Elosua and Geisinger (2016), and is contemplated in the introduction to section 2.11 of the CET-R Furthermore, it was insisted on with the proposal of Gómez (2019) to reconsider the assessment of the label «No information is provided» in the CET-R. In fact, writing in the manual of a test the statement «test X can be used to evaluate characteristic Y» requires different backing than writing the statement «test X can be used to detect a person's difficulties in characteristic Y, to design an intervention plan aimed at improvement of this characteristic, and to monitor its evolution in the clinical, educational, social, and legal fields». The difference lies in the fact that the first statement leaves it to the responsibility of the test user the concrete use that will be made of that assessment and, therefore, the responsibility to support such use (AERA et al., 2014, standard 9.4). On the other hand, the second statement explicitly promotes the test for various specific uses, so the responsibility to support each specific use lies with the publisher (AERA et al., 2014, standards 5.0 and 7.1).

TABLE 3
SUMMARY OF SCORES OF TESTS ASSESSED IN THE EIGHTH EDITION

Characteristic	BADyG						BRIEF	CELF	MCMIIV	PECO	TONI	Average 2019
	i	E1-r	E2-r	E3-r	M-r	S-r	P	5	IV		4	(previous)
Development: Materials and documentation	4.5	4.5	4.5	4.5	4.5	4.5	4.8	4	4.3	3	4.5	4.3(4.3)
Development: Theoretical foundation	5	5	5	5	5	5	4.5	4	4	4	5	4.7(4.1)
Development: Adaptation	-	-	-	-	-	-	5	5	4.5	-	-	4.8(4.3)
Development: Item analysis	4	4	4	4	4	4	4	2	2	4	4	3.6(3.8)
Validity: content	4	4.5	5	5	5	4.5	3.3	3.5	3	3	3.5	4.0(3.8)
Validity: relationship with other variables	3.5	3.4	2.7	3.7	3.7	3	3.3	3.4	3.4	2.6	3.9	3.3(3.6)
Validity: internal structure	2.5	3	3	3	3	3	2.5	-	-	2.5	5	3.1(3.7)
Validity: DIF analysis	-	-	-	-	-	-	-	-	-	-	5	-
Reliability: equivalence	-	-	-	-	-	-	-	-	-	-	3	-
Reliability: internal consistency	4	4	5	5	5	4	5	4.5	4	3.5	3.5	4.3(4.2)
Reliability: stability	-	-	-	-	-	-	4	2.5	3	-	3	3.1(3.5)
Reliability: TRI	-	-	-	-	-	-	-	-	-	-	-	-
Inter-judge reliability	-	-	-	-	-	-	-	-	-	-	-	-
Scales and interpretation of scores	4.7	4.3	4.7	4.3	4.7	4.7	3.7	3.3	4	2.3	4.3	4.1(4.1)

Note. 1: Inadequate, 2: Adequate with deficiencies, 3: Adequate, 4: Good, 5: Excellent, -: Not applicable or No data provided. In brackets: Average score from the averages of the editions from 2010 to 2018.

Another improvement would be achieved if, before presenting the validity results, the concrete hypotheses to be tested were clearly specified as well as which results would be considered as favorable evidence, taking into account the theoretical foundation, the results obtained in previous research, and the intended uses of the test (e.g., Ziegler, 2014). In this sense, the hypothesis-results-conclusions chain should be clear in relation to all factor loadings, all correlation coefficients, and all effect sizes published, even if they are presented within tables or as previously published results. This does not prevent the incorporation of numerous variables in a few hypotheses, as is the case in factorial analysis or in the design of multitrait-multimethod matrices; on the contrary, this would be a highly recommendable format.

The third opportunity for improvement in the area of validity would be to incorporate more information on equivalence and fairness in the use of the tests. Consistently, the previous reviews have encouraged this by increasing the publication of studies on differential item functioning (DIF; e.g., Fonseca-Pedrero & Muñiz, 2017; Gómez, 2019; Hidalgo & Hernández, 2019). However, in the present edition use of DIF is maintained at a similar or lower level, since only one DIF analysis has been provided, the one related to the TONI-4 test. For our part, we evaluate the cost of doing this type of test, including the risk of overestimating the presence of DIF. Therefore, we believe that the time has come to recommend a more affordable, yet fundamental, step to give fairness the importance it is currently recognized both in society and in the regulatory texts (AERA et al., 2014; COP, 2015a, 2015b; Asociación Española de Normalización y Certificación [Spanish Association for Standardization and Certification, AENOR], 2013). Specifically, we suggest including in the manuals a specific section dedicated to providing information on the flexibility of the test to address the functional, linguistic, neurological, or social diversity of potentially assessable persons. This information can be collected through DIF analysis, but only when the groups are large and the hypotheses well defined. In contrast, it is much more feasible to obtain relevant information by consulting specialists and members of minority groups (AERA et al., 2014, standard 3.11). This information can and should be obtained and shared in a rigorous manner (Levitt et al., 2018). This rigor should be extended to all evidence obtained through the use of qualitative methods, including evidence obtained during the development and adaptation of a test and especially the collection of evidence related to the content of the test and the collection of evidence related to cut-off points for the interpretation of scores related to the criterion. In our opinion, as a minimum requirement, it should be specified separately how many people participated and their qualification to do so, the strategies used to obtain the information, the data obtained, whether verbal or numerical, and the conclusions derived from them, in order to facilitate the formation of the reader's own judgment about the quality of those conclusions.

Finally, with respect to reliability, the internal consistency studies included in the manuals that were evaluated mostly received a rating between good and excellent, which means that high reliability coefficients were published obtained with sufficiently large samples. However, the large number of boxes without information that are observed in this section of Table 3 are worthy of mention. Although not all designs for studying reliability are applicable to all tests, both in the CET-R (section 2.12.1) and in previous reviews (e.g., Hernández et al., 2015) the possibility of providing several reliability coefficients for each scale or subscale and subpopulation is suggested as a good practice. For our part, we believe that this could be made concrete by each test providing reliability data obtained with at least two types of designs from those contemplated in the CET-R or, alternatively, giving clear explanations as to why other sources of random variation beyond those derived from the coherence between the test items are not a concern. Even without using any other design than that of internal consistency, it would be very appropriate to recognize that not all the scores in the same test have the same reliability, which would involve the publication of data of relative reliability such as those that can be obtained using item response theory. With this suggestion we again add to the recommendations made in previous reviews (e.g., Gómez, 2019).

Innovative Contributions of the University to the Test Review Process

In this edition, we would like to add our gratitude to the collaboration of the test publishers who offer their help and experience during the test review process, year after year, and the support of the COP Test Commission during the whole process. All the more so because our teaching project has involved greater economic support and adjustment to academic times on their part. We would also like to reflect on the process of prioritizing the tests to be evaluated based on three pieces of information. The first is the ordered list of the 25 most used tests by the associated members [of the COP] according to a recent survey (Muñiz et al., 2020). The second is the list of the 84 tests with reports published on the COP website since 2010. A comparison of the two lists shows that they only coincide in 12 cases and are not always the most used tests. The third is the list of the 91 tests that have been evaluated in this same period by our students in their coursework with varying degrees of success, and this list can be accessed by contacting the corresponding author. These are tests selected from those available in the faculty test library collection, which, in turn, makes acquisitions in response to requests from professors of the psychology faculty. More than a quarter of the 91 coincide with the tests evaluated by the COP, but it is interesting to note the presence of nine tests that are also on the list of the most used which do not have an official published report. Another interesting fact is that one of the tests that is on the list of the most used and that has not been evaluated until now is a test that has not been commercialized. All of this leads one to think that, if participation were opened up to new groups with interests in this process, it would be likely to achieve greater coverage of

the real needs for information on the part of the profession and, furthermore, would be a new step towards diversifying the voices that participate in the review process.

Regarding the performance of our students, the last column of Table 3 shows the comparison of the average scores of the eighth edition with the average scores of the seven previous editions as a whole. Only the 10 characteristics for which it has been possible to obtain average data in all the available reviews are included. Although when looking at the data point by point, a number of divergences are found, which balance out because some are high and others low, the two series of data have the same average value of 3.9 points and the correlation coefficient between them is 0.90; in other words, they are very comparable evaluations overall. Be that as it may, we must not lose sight of the fact that the differences could be attributed to the reviewer, but also to the tests that were evaluated in this edition.

More informative is the comparison of the reports of Review 1 (specialist) with those of Review 2 (students) on the same test. The median of the correlation coefficients between the scores given by Reviewer 1 and Reviewer 2 on the questions in which all the reviewers gave valid scores to all the tests evaluated was 0.67, a moderate value and very similar to that published by Ponsoda and Hontangas (2013), which was 0.61. In fact, a similar value or even lower would be expected in a peer review process where discrepancies are consubstantial, as noted by Fonseca-Pedrero and Muñiz (2017).

In the open-ended questions, the winning student teams wrote much longer texts (between 2000 and 4400 words) than the professionals (between 800 and 3000 words). Comparative reading showed that the students' texts may have been argued in more detail, but they were also more redundant and more dependent on the way of presenting the information used in the textbooks and test manuals. We, the professors, attribute these results to several reasons. Firstly, it may be that students feel the need to include theoretical support that gives them the opportunity to express their opinion. Secondly, their previous training has been based largely on the reading of educational manuals and scientific articles, so it may be difficult for them to deal with texts developed for the marketing of a product, even if it has a scientific basis. On the other hand, experts would have more resources to interpret and evaluate these materials. Another explanation may come from the teachers' comments on the drafts, since students were basically asked to reconsider the inconsistencies of their comments, to incorporate information, and/or to further develop their arguments.

Related to the above, the winning teams submitted comments very much in line with the instructions and headings of the CET-R. This is not surprising because, as we have said, our Psychometrics classes are aligned with these instructions and headings, and the students' questions about them were answered by face-to-face tutoring. However, specialists have also commented that some of the questions in the CET-R are difficult to answer. Therefore, we believe that it would be very helpful to implement the proposal of Fonseca-Pedrero and Muñiz (2017) to lower this barrier by providing more technical information and creating tutorials on how to fill out the CET-R.

Still in the same order of things, we have detected a wide margin for continuing to develop a shared language, both technically and in terms of inclusiveness. As for the technical language, we align ourselves with the opinion expressed in previous works that the CET-R is a good guide for the construction, editing, and use of tests (Elosua & Geisinger, 2016; Muñiz & Fonseca-Pedrero, 2019) and, therefore, we suggest to the publishing houses that when writing the manuals they should use the psychometric language as it is expressed in the CET-R as much as possible. Nothing could be further from our intention than to restrict the presentation of new evidence in support of the uses of a test; on the contrary, it is very welcome. But for the presentation of more classical tests, we propose as a reference the psychometric language of the CET-R because this is a consensus synthesis of some of the most widely accepted normative texts such as the criteria of the EFPA, the standards of the AERA and the APA, of the International Test Commission, and ISO-10667 (Hernández et al., 2016) and also because homogenization would greatly facilitate the sharing of the material with the rest of the profession and especially with beginners. And in relation to the use of inclusive language, on the one hand, in the manuals of the tests that we have evaluated, generic male language is widely used when referring to people and, on the other hand, in our university it is considered good practice to positively value the use of inclusive language. Thus, paradoxically, our students were penalized in their writings for a linguistic practice that is used in the manuals they were evaluating. We believe that this is a good moment to propose to the publishing houses that they join us in setting an example, aligning themselves with the policy of the Spanish Psychological Association regarding the use of inclusive language.

Finally, on the educational side, it is worth noting that the course had a retention rate of over 99% and a success rate of 93%, results that are within the range of those obtained in recent years (UAB, n.d.-a) and that we consider very satisfactory. The few students who participated in the satisfaction survey reflected polarized opinions. Among the negative ones, it was highlighted that a lot of time is dedicated to carrying out the work and this competes with the time dedicated to the explanation and assimilation of theoretical concepts. Among the positive ones, it predominated that the project implies the strengthening of conceptual learning through the real application of theory and the approach to the world of work or the profession. This polarization was reflected in practically the same terms in the teachers' assessments.

CONCLUSIONS AND RECOMMENDATIONS

The first objective of this work was to assess the quality of 11 tests submitted to the eighth edition of the test review in Spain by applying the evaluation model agreed by the COP Test

Commission and reflected in the questionnaire CET-R. Our conclusions are that the documentation accompanying these tests presents more than adequate data to support the development and adaptation of the test as well as the norming samples. Excellent data are also provided regarding the internal consistency of the test, but the presence of other reliability evidence should be increased. We have detected a greater margin of improvement in terms of the provision of evidence to support the validity of the test for each and every one of the proposed uses. Our proposals for improvement are the explicit association of the validity evidence with each of the proposed uses, the prior specification of the hypotheses to be tested, the development of tests in favor of fairness in the treatment of the various evaluable people, and the reporting of the qualitative methodology based on updated standards.

These conclusions have derivatives that affect the structure and assessment of the questions of the CET-R. In this sense, we recommend (a) incorporating a question about the intended uses of the test in the description section and structuring the assessment of validity according to these uses, (b) giving two evaluation options to characterize the missing information, which would be rated either as *Not relevant* or as *Relevant, but no information is provided*, (c) to evaluate in a structured way the validity evidence obtained with a qualitative methodology, considering the method and the results separately, and (d) to promote the fair application of the tests by carrying out in a structured way the evaluation of the data presented about accommodations.

Our second objective was to assess the innovative contribution of the university to two aspects of the review process. Regarding the tests that are submitted for evaluation, our conclusion is that the coverage of the information needs of the profession could be extended if the opinion of other entities besides the COP Test Commission were incorporated into the prioritization process. Furthermore, we have provided data on the extension of the voices represented in the review, by offering this opportunity to students under the tutoring of the teaching staff. Our conclusions are that at a quantitative level there has been a great similarity with previous editions and, at a narrative level, our students have written longer texts that are more adjusted to the instructions of the CET-R, although they did not always provide the most solid arguments because their texts were very dependent on the subject manual and the evaluations expressed in the documentation they were assessing. This has led us to emphasize the usefulness of expanding the development of shared language among different groups with interests in the use of tests. Regarding the educational function, we conclude that it was very effective, since our students were successful in the subject in 93% of cases, although we also observed polarized opinions between the amount of effort the project involved and how motivating it was.

Our conclusion from this experience is that the incorporation of tutored students into the test review process has been costly in terms of the materials and time needed to develop it; it has been variable in terms of its motivational potential; and it has been satisfactory in terms of the students' academic success, the professionalism of the reports they have written, and the ideas they have contributed to develop test manuals that are suitable for beginning professionals.

CONFLICT OF INTEREST

There was no conflict of interest.

ACKNOWLEDGMENTS

We would like to thank Dr. Laura Gómez and Prof. Dr. José Muñiz for their support throughout the process. To the test publishers for the contribution of six copies of each test, to the Faculty of Psychology and to the Humanities Library of the UAB for the purchase of the missing copies to complete the necessary material to cover all the class groups. To the *Espai de Suport i Innovació Docent of the Faculty of Psychology* of the UAB for their support in arranging access to the materials. To Remei Prat, Elena Ripollés, Salomé Tàrrega, and Joan Pons for their contribution to the development of the project as part of the teaching team in previous courses.

REFERENCES

AENOR. (2013). *Prestación de servicios de evaluación. Procedimientos y métodos para la evaluación de personas en entornos laborales y organizacionales. Parte 2: Deberes del proveedor de servicios [Provision of evaluation services. Procedures and methods for the assessment of people in working and organizational environments. Part 2: Tasks of the service provider]*. (UNE-ISO 10667-2:2013) https://www.aenor.com/normas-y-libros/buscador-de-normas/UNE?c=N0051261

AERA, *American Psychological Association* [APA], & *National Council on Measurement in Education* [NCME]. (2014). *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Association.

COP. (2015a). *Directrices internacionales para el uso de los tests [International Guidelines for the use of Tests]*. https://www.cop.es/index.php?page=directrices-internacionales

COP. (2015b). *Principios éticos de la evaluación en psicología [Ethical principles of evaluation in psychology]*. https://www.cop.es/index.php?page=principios-eticos

Doval, E., Viladrich, C., Aliaga, J., Espelt, A., García-Rueda, R., Penelo, E., Prat, R. & Tárrega, S. (2013, 3- 6th September). *Las asignaturas de contenido psicométrico en la UAB: saber y oficio [The subjects of psychometric content at the UAB: knowledge and trade* [communication]. XIII Congreso de la Asociación Española de Metodología de las Ciencias del Comportamiento [XIII Congress of the Spanish Association of Methodology of Behavioral Sciences]. La Laguna, Spain.

Elosua, P., & Geisinger, K. F. (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*(2), 82–88.

Espelt, A., Viladrich, C., Doval, E., Penelo, E., & Aliaga, J. (2016, 5-7th July). *Relació entre l'adherència al funcionament de l'assignatura de Psicometria i la qualificació final dels estudiants [Relationship between adherence to the functioning of the of Psychometrics and the final qualification of the students]* [poster]. IX Congreso Internacional de Docencia e Innovación Universitaria [IX International Congress on University Teaching and Innovation] (CIDUI). Barcelona, Spain.

Evers, A., Muñiz, J., Hagemeister, C., Hstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Evaluación de la calidad de los tests: revisión del modelo de evaluación de la EFPA [Assessing the quality of tests: Revision of the EFPA review model]. *Psicothema, 25*(3), 283–291. https://doi.org/10.7334/psicothema2013.97

Fonseca-Pedrero, E., & Muñiz, J. (2017). Quinta evaluación de Tests editados en España: mirando hacia atrás, construyendo el futuro [Fifth review of tests published in Spain: Looking back, building the future]. *Papeles del Psicólogo, 38*(3), 161–168. https://doi.org/10.23923/pap.psicol2017.2844

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

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G., & Elosua, P. (2016). Revisión del modelos 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*, 161–168.

Hernández, A., Tomás, I., Ferreres, A., & Lloret, S. (2015). Evaluación de tests editados en España [Evaluation of tests published in Spain]. *Papeles del Psicólogo, 36*(1), 1–8.

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

International Test Commission. (2018). *ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations.* https://www.intestcom.org/page/31

Levitt, H.M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R. & Suárez-Orozco, C. (2018). Journal Article Reporting Standards for Qualitative Primary, Qualitative Meta-Analytic, and Mixed Methods Research in Psychology: The APA Publications and Communications Board Task Force Report. *American Psychologist, 73*(1). 26-46. https://dx.doi.org/10.1037/amp0000151

Muñiz, J., & Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test. [Ten steps for test development]. *Psicothema, 31*, 7–16. https://doi.org/10.7334/psicothema2018.291

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

Muñiz, J., Fernández-Hermida, J. R., Fonseca-Pedrero, E., Campillo-Álvarez, Á., & Peña-Suárez, E. (2011). Evaluación de tests editados en España [Review of tests published in Spain]. *Papeles del Psicólogo, 32*(2), 113–128.

Ponsoda, V., & Hontangas, P. (2013). Segunda evaluación de tests editados en España [Second evaluation of tests published in Spain]. *Papeles del Psicólogo, 34*(2), 82–90.

Redondo-Corcobado, P., & Fuentes, J. L. (2020). La investigación sobre el Aprendizaje-Servicio en la producción científica española: una revisión sistemática [Research on Service-Learning in the Spanish scientific production: A systematic review]. *Revista Complutense de Educacion, 31*(1), 69–83. https://doi.org/10.5209/rced.61836_

UAB. (n.d.-a). *Seguiment de titulacions: Grau en Psicologia, Psicometria. [Follow-up of qualifications: Degree in Psychology, Psychometrics].* http://siq.uab.cat/siq_public/titulacio/2502443/assignatura/102569

UAB. (n.d.-b). *Enquesta de satisfacció d'assignatura de la UAB. [Satisfaction survey of subjects at the UAB].* https://www.uab.cat/doc/QuestionariEnquestaAssignatures

Vermeulen, K. (2019). English version of the COTAN review System. *Testing International, 41*, 8.

Viladrich, C., Doval, E., & Penelo, E. (2014, 23-25th July). *Student versus expert test reviews: What can we learn from them?* [Communication from the symposium Viladrich, C. (presidency) Symposium Tests review as a tool to enhancing testing practices]. VI European Congress of Methodology. Utrecht, Holland.

Viladrich, C., Doval, E., Penelo, E., Aliaga, E., Espelt, A., García-Rueda, R., & Angulo-Brunet, A. (2019). *Avaluació de tests psicológics*. [*Evaluation of psychological tests*]. Assignatures i pràctiques ApS. http://pagines.uab.cat/aps/ca/content/assignatures-i-pràctiques-aps

Viladrich, C., Doval, E., Penelo, E., Aliaga, E., Espelt, A., García-Rueda, R., & Angulo-Brunet, A. (2021, 21-23th July). *Eighth edition of the Spanish evaluation of test quality: A service-learning experience.* [abstract accepted at the symposium Hernández, A. & Muñiz, J. (presidency) Improving tests and testing practices: international and multi-stakeholder perspectives]. IX European Congress of Methodology. Valencia, Spain.

Viladrich, C., Doval, E., Aliaga, J., Espelt, A., García-Rueda, R., Penelo, E., Tárrega, S., Ripollès, E., & Prat, R. (2014). Aprendices de certificadores en psicometría: una experiencia ABP con grupos grandes [Certifier Apprentices in Psychometrics: A PBL Experience with Large Groups]. *Revista del CIDUI, 2*,1–9.

Ziegler, M. (2014). Stop and state your intentions! Let's not forget the ABC of test construction. *European Journal of Psychological Assessment, 30*(4), 239–242. https://doi.org/10.1027/1015-5759/a000228