SECOND EVALUATION OF TESTS PUBLISHED IN SPAIN

Vicente Ponsoda[1] and Pedro Hontangas[2]
[1]Autonomous University of Madrid. [2]University of Valencia

This article describes the results of the second evaluation of psychological tests published in Spain. The Official Association of Psychologists Testing Committee decided that 12 tests, selected mainly because they were innovative and widely used, should be assessed. Each test was evaluated by two experts. As in the first evaluation (Muñiz, Fernández-Hermida, Fonseca-Pedrero, Campillo-Álvarez and Peña-Suárez, 2011), the evaluators' work was done by responding to the questions on the Questionnaire for Test Evaluation (Prieto and Muñiz, 2000), which adapts the model prepared by the European Federation of Professional Psychologist Associations to the Spanish context. Each test was evaluated for quality of materials and documentation, score reliability, coverage of validation studies, quality of scales, etc. The reviewers also reported on the suitability of the instrument and the evaluation process and made suggestions that might be useful to improve future assessments.

***Keywords**: Tests, Use of tests, Test evaluation, Psychometry.*

El artículo describe los resultados de la segunda evaluación de tests psicológicos editados en España. La Comisión de Tests del Colegio Oficial de Psicólogos decidió que se evaluasen 12 tests, seleccionados principalmente por su novedad y amplio uso. Cada test ha sido evaluado por dos expertos. Al igual que en la primera evaluación (Muñiz, Fernández-Hermida, Fonseca-Pedrero, Campillo-Álvarez y Peña-Suárez, 2011), los evaluadores hacían su trabajo respondiendo a las preguntas del Cuestionario para la Evaluación de los Tests (Prieto y Muñiz, 2000), que adapta al contexto español el modelo elaborado por la Federación Europea de Asociaciones de Psicólogos Profesionales. De cada test se ha evaluado la calidad de los materiales y documentación, la fiabilidad de sus puntuaciones, la cobertura de los estudios de validación, la calidad de los baremos, etc. Los revisores informaron también acerca de la idoneidad del instrumento y proceso seguido en la evaluación. Se aportan sugerencias que pudieran ser útiles para mejorar las evaluaciones futuras.

***Palabras clave:** Tests, Uso de los tests, Evaluación de tests, Psicometría.*

Psychological tests are tools widely used by psychologists to make decisions with considerable social impact on the various fields of psychology, e.g., educational, clinical, social, organizational and legal. It is no surprise, then, that in several European countries their tests are systematically evaluated. This is the case in the United Kingdom, Germany, Norway, and especially, the Netherlands. Evers, Sijtsma, Lucassen and Meijer (2010) report on the main characteristics of the Dutch evaluation process. Over 40 years ago, the Dutch Psychologists Association Committee on Testing started up a first evaluation of the quality of their tests, which led to a first book of evaluations, published in 1969. Five more have followed this first publication, the latest in 2009. From 1982 to 2010, 878 tests were reviewed, which are practically all those published. Therefore, in the Netherlands, anyone interested in giving a test can find an independent, rigorous evaluation of its quality and properties which are practically all of those published. Therefore, in the

Netherlands, anyone who is interested in giving a test can find an independent, rigorous evaluation of its quality and properties.

Muñiz and Fernández-Hermida (2010) showed that the opinion of Spanish psychologists on the use of tests is clearly positive. On a scale of 1 ("completely disagree") to 5 ("completely agree"), the mean on the item "Used properly, tests are of great help to the psychologist" was 4.41. They also agree (mean = 4.13) with the item "The Official Association of Psychologists should exert a more active role in regulating and improving the use of tests." In the same study, the agreement above turns into slight disagreement (mean = 2.71) when the sentence is "Professionals have sufficient information (independent reviews, research, documentation, etc.) on the quality of tests published in our country." These results encouraged the Official Association of Psychologists (COP), through

Correspondence: Vicente Ponsoda. *Facultad de Psicología. Universidad Autónoma de Madrid. C/ Iván Pavlov 6. 28049 Madrid. España. E-mail: Vicente.ponsoda@uam.es*

its Committee on Testing (CT), to begin systematic review of the tests published in Spain, in the wake of evaluations other countries have been making. In 2010, the process began and ten tests were reviewed. The main results were published in this journal (Muñiz, Fernández-Hermida, Fonseca-Pedrero, Campillo-Álvarez and Peña-Suarez, 2011). The complete review of each test is available on the COP Web (http://cop.es/8). The experience was considered positive by the CT, which decided to make a second evaluation. This article describes the main results.

EVALUATION PROCESS

The process followed in this second evaluation coincides basically with the first, with some differences mentioned as appropriate. In the first place, the CT decided that tests should be reviewed, and that the number of tests should be 12. Eleven of these tests are marketed by the test editors who make up the CT (three EOS, three Pearson and five TEA). The twelfth test is the EPV-R scale (Echeburúa, Amor, Loinaz and Corral, 2010), which is not yet on the market, and is being applied for detecting high risk of severe domestic violence in many police stations. The list of the 12 tests is shown in Table 1. The next step was to decide who would be the coordinator. At the proposal of the CT, the first author of this study agreed to do it.

Then the search for reviewers began. As in the first evaluation, it was thought that it would be best for one evaluator to be a better expert in psychometrics and the other in the subject variables the test measured. Special care was given the following three matters: In the first place, it was attempted to choose an expert in psychometry who was interested in the content and knew more about it to evaluate the test. Something similar was done with the content experts. Of the experts in a particular content, the one chosen was the one who had publications most closely related to the tests. In the second place, the coordinator decided not to recur to any of the reviewers who participated in the first evaluation so that little by little a bank of reviewers could be generated, since the CT intends to continue evaluating tests in coming years. Finally, it was attempted to choose reviewers who were not directly related to the authors of the tests. In fact, on the letter of invitation, they were told not to accept participation in the review if they doubted whether they could make an objective assessment.

A first list with two possible reviewers for each test was analyzed by another two members of the CT. They found that one reviewer was not appropriate and the list was modified. In continuation, the coordinator invited them to participate. Of the 24 invitations, there were only two refusals. In one case, for very understandable personal reasons, and in the second, because it was felt that the evaluator's assessment might not be objective.

The selection of reviewers, which is a transcendental point, is one of the aspects that turned out to be the most

TABLE 1
LIST OF TESTS EVALUATED

BAI	Beck Inventory of Anxiety
BAS-II	Intellectual aptitude scales (British Ability Scales)
BDI-II	Beck Depression Inventory II
CEAM	Motivated Strategies for Learning Questionnaire
CompeTEA	Competency Assessment Questionnaire
EPV-R	Severe Domestic Violence Risk Prediction Scale
ESCOLA	Reading Awareness Scale
ESPERI	Questionnaire for Detection of Behavior Disorders in Children and Teenagers
Merrill-Palmer-R	Revised Merrill-Palmer Development Scales
PAI	Personality Assessment Inventory
RIAS/RIST	Reynolds Intellectual Assessment Scales/Reynolds Brief Intelligence Test
WNV	Wechsler Nonverbal Intellectual *Aptitude Scale*

TABLE 2
REVIEWERS WHO EVALUATED THE TESTS

Reviewer	Affiliation
Francisco José Abad García	Autonomous Univ. Madrid
Jesús Alonso Tapia	Autonomous Univ. Madrid
Jesús Alvarado Izquierdo	Complutense Univ. Madrid
Juan Antonio Amador Campos	Univ. Barcelona
Ramón Arce Fernández	Univ. Santiago de Compostela
Roberto Colom Marañón	Autonomous Univ. Madrid
Pere Joan Ferrando Piera	Univ. Rovira y Virgili
Eduardo García Cueto	Univ. Oviedo
Paula Elosua Oliden	Basque Country Univ.
Pedro Hontangas Beltrán	Univ. Valencia
Fernando Jiménez Gómez	Univ. Salamanca
Antonio Lobo Satué	Univ. Zaragoza
Raúl López Antón	Univ. Zaragoza
Ramón López Sánchez	Complutense Univ. Madrid
Antonio Maldonado Rico	Autonomous Univ. Madrid
María Rosario Martínez Arias	Complutense Univ. Madrid
Julio Olea Díaz	Autonomous Univ. Madrid
José Olivares Rodríguez	Univ. Murcia
José Luis Padilla García	Univ. Granada
Lilisbeth Perestelo Pérez	Canary Islands Health Service
Gerardo Prieto Adánez	Univ. Salamanca
María Ángeles Quiroga Estévez	Complutense Univ. Madrid
Jordi Renom Pinsach	Univ. Barcelona
Jesús Salgado Velo	Univ. Santiago de Compostela
Carme Viladrich Segués	Autonomous Univ. Barcelona

satisfactory. As shown in Table 2, their scientific quality is hard to improve on and so is their involvement and fine work in the various stages of the process. The coordinator would like to take advantage of these lines to express his profound appreciation to them all.

The editors made two complete copies of each test available to the COP. The COP sent one to each reviewer. With regard to the EPV-R test, which is not on the market, the coordinator explained to the main author that his test was going to be reviewed and asked him to indicate the minimum documents the reviewers could work on. His answer was positive and very cooperative, and thus the documents the two reviewers would have to evaluate were determined.

As in the first evaluation, evaluations were done by responding to the Questionnaire for Test Evaluation (CET), which adapts the European Federation of Professional Psychologist Associations test evaluation model to the Spanish context (Prieto and Muñiz, 2000).

The following lines provide a brief description of the CET, since the rest of the article makes constant reference to its characteristics. It has three sections. The first (General Description) consists of 31 questions on the name of the test, authors, date of publication, description of what it measures, areas of application, item format, administration, price, etc. The second section (Evaluation of Characteristics) consists, in turn, of several subsections and contains a total of 35 questions. The questions in the first subsection have to do with the quality of the materials, documentation and instructions, the theoretical basis, ease of administration, quality of the adaptation process and formal analysis and psychometrics of the items. The questions in the second subsection (validity) evaluate content validity, construct, predictability and bias of items. The third subsection (reliability) includes questions on equivalence, internal consistency and stability indicators. The last subsection evaluates the quality of the scales. In the third and last section (overall evaluation) a list of the test's strong points and weak points and two summary tables have to be filled in. The first has to be completed based on 31 questions in the General Description Section and the second requires evaluation of 12 characteristics (listed in the first column of Table 3) that summarize the evaluations provided in the Evaluation of Characteristics Section. Of the 68 questions on the CET, 25 are quantitative and have to be answered on a 5-point Likert-type scale (from 1 "inadequate" to 5 "excellent"). On each question, the concrete meaning of "excellent" is defined. The rest are open questions.

Each reviewer was told that his main task was to apply the CET to the test that he had been assigned and he was then informed briefly about the various stages of the process. When the evaluations were received, it was the coordinator's task to generate a combined evaluation based on the two received.

The last step was to send the combined evaluations to the editors and authors of the noncommercial test to get their opinions and comments. The answers varied appreciably in length and in agreement with what was shown in the evaluation sent to them. In our opinion, the participation of the editor/author is a very important point in ensuring the quality of the final review. Based on the answers received, the coordinator modified the evaluations, when he considered it appropriate and generated the final evaluations. The last step was to present them to the CT for its knowledge and approval before publishing them.

RESULTS

As mentioned above, the CET requires qualitative and quantitative evaluations. Not all of the questions can be scored. For example, a question that asks about the quality of the adaptation, when not all of the tests evaluated are adaptations. In others, the manual may not offer information on what is asked. To acquire evidence of reliability among the evaluators, the correlation of the scores given by the two evaluators of each test on the questions (5) on which all the tests evaluated have scores was calculated. The median of the 12 correlations, the same number as tests evaluated, is 0.61.

The main results are shown in Table 3, which contains the evaluations of the 12 tests on each of the 12 characteristics included in the evaluation summary table. The second-to-the-last column contains the means of the evaluations for each characteristic (minimum of 1 and maximum of 5). The highest mean is 5 and corresponds to the characteristic "Analysis of bias", although it was only found for one test. In the first evaluation, none of the ten tests was evaluated for this characteristic. The next best means are 4.32 and 4.29 and correspond to "Quality of materials and documentation" and "Spanish adaptation". The lowest is 3.40 and corresponds to the characteristic "Reliability reporting stability". The neutral point of the response scale is 3 ("Adequate"). Even the lowest mean surpasses this neutral point. The two next worst are 3.50 and 3.55 which correspond to “Content validity” and “Predictive validity”, respectively.

In the first evaluation, the characteristics that were evaluated best, with 4.5 and 4.35, were the “Reliability using equivalence indicators” and “Quality of materials and documentation” respectively. The worst one, with a mean of 3.5, was the indicator of “Reliability reporting stability” as was the case in the second evaluation. Comparing the results of the last two columns, it may be observed that “Content validity” was clearly worse in the second evaluation, while “Construct validity” was worse in the first.

The total mean of the second evaluation was 4.02, slightly above the mean for the first evaluation (3.96). Eliminating the characteristic “Analysis of bias,” which was only evaluated on one of the 22 tests in both evaluations, from the second evaluation, it would be 3.93. The mean evaluation is very close to 4, rating which corresponds to the label “Good” on the response scale. Evers et al. (2010) report on the means of the evaluations done by the Dutch reviewers. The response scale in this case is for

three categories: 1 (“insufficient”), 2 (“sufficient”) and 3 (“Good”). The mean of the last 540 tests reviewed is 2.03, very near the neutral point on the response scale (2), when the Spanish means are clearly above the neutral point on the corresponding scale. There are several differences between the Dutch and Spanish reviewing systems, so it is not easy to explain the differences found. One important difference is that the Dutch evaluate seven criteria, while the Spanish evaluations shown in Table 3 do so for 12, which questions that the meaning of the scores is really the same. In Table 3, a dash shows absence of information or that it is not applicable. It is possible that some of the dashes that show absence of information should really be a low score. In this case, of course, the means would be lower.

COMMENTS ON THE CET AND THE TEST EVALUATION PROCESS

Concerning the CET

Muñiz et al. (2011) suggest that it would be better to modify the CET in the light of the changes introduced in the new European evaluation model (Evers et al., 2010) and improve the questions that had become problematic to a certain extent in the first evaluation. Based on the reviewers’ comments in the second evaluation, we can make some additional suggestions.

In each of the 25 quantitative questions on the CET, we found the variance of the two evaluations made by the

TABLE 3
SUMMARY OF RATINGS OF TESTS EVALUATED

	Tests Evaluated												Mean 2012	Mean 2011
Characteristics	BAI	BAS-II	BDI-II	CEAM	Compe TEA	EPV-R	ESCOLA	ESPERI Palmer-R	Merrill	PAI	RIAS RIST	WNV		
Quality of materials and documentation	4.5	4.5	5	3.5	4.5	-	3.5	3	5	4.5	4.5	5	4.32	4.35
Theoretical basis	4	5	5	4	4	3	3	3	4	5	4	4	4.00	4.20
Spanish adaptation	4	5	4	-	-	-	-	-	5	4	4	4	4.29	4.25
Item analysis	4	5	4	4	4	3	4	3	4	4	4	-	3.91	3.58
Content validity	3	4	4	-	3.5	2.5	3.5	2	4	4	4	4	3.50	4.25
Construct validity	4.5	5	5	4	4	2	2.5	4	4.5	4.5	5	5	4.17	3.60
Analysis of bias	-	5											5.00	-
Predictive validity	4	4											3.55	3.57
Reliability: equivalence	-	-	-	-	-	-							4.00	4.50
Reliability: internal consistency	4.5	4	5	4	3	3.5	4	4	4.5	4.5	4.5	4.5	4.17	3.75
Reliability: stability	—	3	-	-	-	-	-	-	4.5	3.5	3-	3	3.40	3.50
Scales	-	4	-	4	4	-	3	4	4	4	4.5	4	3.94	4.00
Overall mean													4.02	3.96

Notes: The scores in the table are given on a 5-point scale: 1= inadequate, 2 adequate but with some insufficiencies, 3=adequate, 4=good, 5=excellent. A dash (-) means no information or not applicable.

two reviewers of each test. In continuation we found the mean of the 12 variances (one per test) as an indicator of the "ambiguity/difficulty of application" of each question. It could be expected that agreement between the two evaluators would be better (and the variance higher) for the problematic questions. Of the 25 questions, the three that had the highest mean variance were 2.9.1 ("Item analysis"), 2.10.1.2 ("Number of experts consulted for content validation") and 2.10.3.2 ("Size of samples for predictive validation"). The disagreement between the reviewers in their evaluation of item analysis could be due to experts in psychometry evaluating specifically the psychometric analysis of the items, while the content experts may have evaluated the quality of the items without paying as much attention to the manual's comments on discrimination, difficulty, details of items eliminated, etc. The questions referring to content validity turned out to be problematic to some extent, since for some evaluators, the information in the manual, when it describes test development in the table of specifications and processes for checking that the items are related to the construct they are intended to measure, is sufficient evidence of content validity. However, for other evaluators, a study of content validity must be a test development follow-up study intended to show proof of whether the test really evaluates the relevant parts of the construct of interest. If there is some disagreement as to what is understood by content validity, it is not surprising that there is also disagreement insofar as the number of experts involved in that validation. Something similar may be said of predictive validation. The line separating convergent construct validity and criterion validity is often very thin. The manuals sometimes present studies in predictive validity that should be considered convergent validation. It is therefore not infrequent for the evaluators to disagree in reporting on sample size.

One matter that future CETs should pay more attention to is what weight to give the original data compared to those found in the adaptation. The CET includes a question for evaluating adaptation quality. The manuals usually offer many results found with the original test and fewer for the adapted test. This is to be expected, as the original test has been available longer and has been applied more often. For example, in the section on construct validity, the manual often offers many studies done with the original test and some done in Spain. What weight should be given to one or the other in construct validity evaluation? Should all the studies be considered in the evaluation or only the latter? This matter may be behind some of the discrepancies observed between evaluators.

In line with what Muñiz et al. suggest (2011) concerning the first evaluation, evaluator instructions should be clearer. Three evaluators of the second evaluation changed the CET choices when they found none that fit what they wanted to say. It would be a good idea to include a set of general instructions indicating that the CET choices should not be changed, and give standards by which the evaluator may or may not add an explanatory note or justification of the quantitative ratings he gives, that questions should not be left unanswered, etc. Perhaps a glossary should be included to define psychometric terms that could lead to doubt in understanding. As Prieto and Muñiz (2000) suggest, the possibility of computerizing CET administration might be considered. It would make answering more uniform, as only one or more than one answer depending on the question, it could provide the meaning of a term by clicking on it, and would calculate the scores on the questions on which it is expected that the score be a mean of scores assigned to other questions, etc.

It might even be considered whether the CET is adequate for all types of tests. The CET contains a section on scales with four questions to evaluate the aids to interpreting scores. However, in some situations (clinical scales, for example), it makes more sense and is more frequent to set cut-off points that enable the score to be classified in one of the groups of interest. Would it be better for the CET to include a section on interpreting scores that would also enable other interpretation strategies to be evaluated as alternatives to the scales? There are tests (Evers et al. 2010) that are not intended to predict outside results and in which predictive validity does not make much sense. On the clinical scales, results are usually the capacity of the test for predicting pertinence to different groups, which do not require correlations to be calculated. The question with which the test's predictive capacity is rated ("Mean of correlations on the test with criteria") does not seem appropriate in this case.

The batteries pose some specific problems too. The CET states the following in a note: "If the test is comprised of subtests with heterogeneous formats and characteristics, fill in the questionnaire for each subtest." It should be emphasized that the work a 50-page manual (e.g., BDI-

II) gives the reviewer is not the same as a battery (e.g., BAS-II, with several long manuals and different tests.) In one case, the reviewer told the coordinator that if he had to follow the instructions in the note, he could not do the review. It should be considered whether that note should be omitted and clearly state what supplementary information is necessary for batteries and on which questions, keeping the amount of work requested of the reviewer as reasonable as possible.

Finally, we could add some further suggestions to the list of subjects that have posed some difficulty in the first and second evaluations: Would it be better to break down construct validity into internal structure and relationships with other variables? Some tests have more items than those administered, since the user has to select those appropriate for each person assessed. It would be advisable to indicate how to proceed in this case (for example, by indicating the number of items available and the maximum number of items it is possible to use on each test). Also on the second evaluation, the question on the CET on "correction procedure" posed some difficulty, since sometimes "computerized" was confused with "done exclusively by the provider." Sometimes the correction procedure is manual, but with no template. Question 2.11.2.1 asks about the sample size. "Several studies with small samples" is not given as a choice. On scales, the quality of norms and the size of the groups is rated, but how is the application of strategies such as "continuous norming" (Zachary & Gorsuch, 1985), which palliates the problem of small normative group sizes, taken into account?

Concerning the reviewing process

Muñiz et al. (2011), weighing the pros and cons of the first evaluation, recognized that the review process posed some doubts, and surely the fact that the different countries follow different procedures has something to do with this. In our country, as described above, test reviewing is done in much the same way as scientific articles are reviewed, but with some differences.

Review of a noncommercial test presents some difficulties of its own, not the least of which is determining which articles/reports the reviewers have to base their evaluation on, as there is no manual. The first noncommercial test reviewed was the EPV-R and it was done in this second evaluation. The coordinator chose to ask for help from the authors, who, by the way, gave it willingly and effectively. The doubt of how to proceed if someone did not remains. It makes sense for the CT to initiate review of any test it considers appropriate and make it public, but it would be advisable to set up a protocol of how to go about it in such cases. It could indicate, for example, whether the author should be informed and his cooperation requested, and who should do it, and also determine what documents the review should be based on, and whether the reasons the test was chosen have to be stated, among other possible contents.

Science sometimes does not match well with business. If the manual reports on many psychometric details, the price of the test is higher because of the effort it takes to make it, and the cost of the manual itself (more pages). What could happen then is that following the recommendations proposed by the reviewers leads to difficulties for its marketing. One possible solution could be that the editors offer the basic information in the manual and the more sophisticated information on the Web. Another example of the same type of thing occurs when the editor does not publish relevant psychometric information in the manual (for example, weights of the items in the factors). The reviewers score negatively because these data are not offered even though the information exists. One possible solution to this is that the editors supply the reviewers with any information available not in the manual that is required by the CET along with the test. This would have to be done safekeeping the confidentiality of this information and the anonymity of the reviewing process. A third example has to do with the consequences of the review: the strong points of the tests reviewed are described, but also its weak points. This information is not available for tests not reviewed, leaving doubt as to whether these are not better off. Evers et al. (2010) state that in the Netherlands, the idea has become established that good practice in the use of tests is to apply the ones which have received good evaluations in the review process. We could not be more in agreement and it is to be hoped that something similar happens in our country when the review process progresses and more and more tests are reviewed.

Returning to what was mentioned at the beginning of the paragraph above, the following question might be asked: Does it make sense to recommend complicated psychometric analyses (invariance tests, differential item functioning studies, indicators of the accuracy of each measure, etc.) and request that the details be provided (for example, information on adjusting the models) that most applied psychologists probably do not understand? What the users will probably ask for are aids in interpreting the ratings. Should the review process incorporate the viewpoint of the professionals more actively? The reviewers are almost all scholars. Are we not biased if we evaluate a test based on articles when the purpose and the target public are others? As mentioned by Elosua (2012), modern psychometry is progressing very quickly and the distance between current developments and those that have been applied in theses, articles, test manuals, etc., by those who are not experts in psychometry is usually considerable. Recently, Ponsoda (2010) coordinated the monographic issue of this journal on "Methodology at the Service of the Psychologist," the purpose of which was to bring some of the modern psychometric developments closer to the professional, who had probably not studied them before, such as the bias of items and tests, confirmatory factor analysis, structural equation models, recent concepts of reliability and validity, new test theories, innovative test and item formats, etc. From our point of view, the new psychometric developments that the reviewers recommend help improve the tests, since new evidence of validity, alternative indicators of test reliability, indicators of the accuracy of individual measurements, of obvious interest for evaluating the individual compared to groups, etc., will result from their application. All of the above is not contradictory to the manual further satisfying the needs and demands of the user and facilitating proper and convenient application of the test. In this same sense, it would be of interest to include the user's viewpoint in the reviewing process, but obviously, it could not be by asking him to respond to the CET. Some procedure would have to be found that reports on his satisfaction with the test, and enables its strong points and weak points to be known from the user's and not the expert's perspective. This seems difficult to integrate in the current reviewing process. Information of this type is found from opinion surveys on tests (Muñiz and Fernández-Hermida, 2000, 2010), although obviously it is not specific to a particular test. With regard to the third question, we do think there is a certain risk in reviewing tests in the same way we do articles, due to the scant experience we all have in reviewing tests. It is the job of the coordinator, in his interaction with the reviewers, to make them see that, in fact, the purpose of the review is not the same as for scientific articles. Therefore, their recommendations must be limited to the subjects that improve the test, provide new proof of validity, etc., and not proposals that could eventually improve the knowledge of psychometric procedures applied or of the construct the test measures.

Some comment might also be made with regard to the role of the coordinator. In the two reviews done till now, each reviewer received the test as a gift from the publisher. The truth is that the coordinator also needs to have the tests for the reports he has to write, to add a third review to the other two if he deems fit, to clarify discrepancies between reviewers and occasionally to check any changes that authors and editors propose in the review that is sent to them before the final evaluation. A possible solution is that the editors and/or the CT provide the center where the coordinator works with the tests to be reviewed which it does not have and cannot purchase.

CONCLUSIONS

In the first place, it should be stressed that the test evaluation process begun in 2010 continues and is becoming consolidated. However, for this consolidation to be more useful, there should be a rapid increase in the number of tests reviewed. To date 22 tests have been reviewed. One of them is a noncommercial test, chosen for its social repercussion. The EPV-R test by Echeburúa et al. (2010) has been included in the protocol followed in many police stations after a report of aggression against a woman to predict the risk of severe domestic violence. Previously, the police decided on what protection was to be given subjectively. With the application of the test, protection is provided as established for the risk level that follows from its application.

The mean level of the tests reviewed is, in absolute terms, good (4 on a scale of 1 to 5), which almost coincides with the first evaluation. Later evaluations will show whether this is the mean quality of the tests published in Spain or whether, as more and more tests are evaluated, the mean changes. We found that only one test in the two evaluations gives detailed information on bias or differential item functioning. Three characteristics are below the mean in both evaluations: "Item analysis," "Predictive validity" and "Reliability understood as stability." In the section on points to improve, it was suggested several times that the manual report on the individual properties of the items and the selection criteria for configuring the final test, more proof of the test's capacity for predicting relevant criteria in their fields of application, and that in new editions of the test, test-retest reliability studies, which inform on score stability, be included.

The reviewers warned about some other deficiencies, such as the following: scant justification for the cut-off points given for interpreting scores, not finding indicators of the accuracy of the score of each individual evaluated, but for the test as a whole, scarcity of differential functioning studies of items and tests, and of studies that provide proof and justify certain expectable uses of the scores. The glass can also be seen as half full. In general, much care has been taken with proper insertion of the construct the test measures in psychological theory. The use of item response theory, which, by the way, provides indicators of the accuracy of each measurement, is observed in several tests. Some differential functioning studies were done and some confirmatory factor analyses and structural equation models were applied. Recent developments were used for the construction of scales that enable many different ones to be found, keeping the total sample size required from becoming too large. In some tests "accommodations," or changes, are included that makes correct interpretation of scores possible when the test is given to individuals with some special characteristic. Several tests provide systems for automatic correction and interpretation of scores, and finally, on some of the tests reviewed, care is taken with representation of normative samples, including non-incidental sampling, and cross-validation is applied to avoid artificially high psychometric indicators. In brief, the good news is that the distance between theory and practice, mentioned in the section above, is becoming shorter.

Evaluation of test quality needs the CET for several reasons. It facilitates the task of the reviewers and test authors and editors, by indicating exactly what is evaluated and how. It enables quantitative and qualitative evaluation of the relevant characteristics to be kept in mind when determining test quality. It facilitates comparison of different test and batch evaluation results. Obviously, this does not exclude the need for revision. In general, the CET could be improved by adding a glossary with terms likely to be interpreted differently and clarifications and examples of problematic questions. Alternatively, it could also include some way to seek reviewer consensus.

The CET is inspired by the European Federation of Professional Psychologist evaluation model, and this model is being modified in depth (Evers et al., 2010). The incorporation of some of these modifications and the answer to the difficulties with its application mentioned in this article and in Muñiz et al (2011) suggest that it would be advisable to think about its modification. Everything would indicate that a new edition of the "standards" (AERA, APA and NCME, 1999) will be published in the coming months. Sharing with Elosua (2012) the idea that we have to bear them in mind, the next appearance of the new ones is another reason for modifying the CET.

Concerning the reviewing process as a whole, our impression is that it works reasonably well. This does not exclude the introduction of some changes as deemed appropriate. The reviewers are given a symbolic amount of 50 Euros. Some reviewers preferred not to take it, since they believe that, as is the case with review of articles, they are tasks that should not be remunerated. Following the model of scientific journals that name the editor and editorial committee for a period of two or three years, it might be considered whether it is also appropriate or not for the coordinator and reviewers to evaluate more than one batch of tests for two or three years, and report on the results of the review at the end of that period.

ACKNOWLEDGMENTS

We would like to thank the members of the Test Commission for their collaboration: José Ramón Fernández Hermida, Ana Hernández Baeza, Miguel Martínez García, Milagros Antón López, Viviana Gutman Mariach, and, especially its President, José Muñiz Fernández, for its continual assistance. We also reiterate our appreciation to the reviewers for their enthusiastic and excellent collaboration.

REFERENCES

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

Echeburúa, E., Amodr, P.J., Loinaz, I. and Corral, P. (2010). Escala de Predicción de Riesgo de Violencia Grave contra la pareja – Revisada (Scale of Prediction of Risk of Severe Domestic Violence-revised) (EPV-R). *Psicothema, 22(4)*, 1054-1060.

Elosua, P. (2012). Tests publicados en España: Usos, costumbres y asignaturas pendientes (Tests published in Spain: Uses, customs and subjects pending). *Papeles del Psicólogo, 33(1),* 12-21.

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

Muñiz, J. and 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. and 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, 32(2),* 113-128.

Ponsoda, V. (2010). Metodología al servicio del psicólogo (Methodology at the service of the psychologist). *Papeles del Psicólogo, 31(1),* 2-6.

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

Zachary, R.A. & Gorsuch, R.L. (1985). Continuous norming: Implications for the WAIS-R. *Journal of Clinical Psychology, 41*, 86-94.