Psicothema

Psicothema 2013, Vol. 25, No. 3, 283-291
doi: 10.7334/psicothema2013.97

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Assessing the quality of tests: Revision of the EFPA review model

Arne Evers[1], José Muñiz[2], Carmen Hagemeister[3], Andreas Høstmælingen[4], Patricia Lindley[5], Anders Sjöberg[6] and Dave Bartram[5]

[1] Dutch Association of Psychologists, [2] Spanish Psychological Association, [3] German Psychological Association, [4] Norwegian Psychological Association, [5] British Psychological Society and [6] Swedish Psychological Association

Abstract

Background: Diverse national and international organizations have been developing projects for many years to improve testing practices. The main goal of this paper is to present the revised model of the European Federation of Psychologists' Associations (EFPA) for the evaluation of the quality of tests. This model aims to provide test users with rigorous information about the theoretical, practical and psychometric characteristics of tests, in order to enhance their use. **Method:** For the revision of the test review model, an EFPA task force was established, consisting of six European experts from different countries, who worked on the update of the previous European model, adapting it to the recent developments in the field of psychological and educational measurement. **Results:** The updated EFPA model provides for the comprehensive evaluation of tests. The first part describes test characteristics exhaustively, and in the second part, a quantitative and narrative evaluation of the most relevant psychometric characteristics of tests is presented. **Conclusions:** A revision of the European model for the description and evaluation of psychological and educational tests is presented. The revised model is analyzed in light of recent developments in the field.

Keywords: Test review, EFPA, Test quality, Test use.

Resumen

Evaluación de la calidad de los tests: revisión del modelo de evaluación de la EFPA. **Antecedentes:** el objetivo de este trabajo es presentar una revisión del modelo de la Federación Europea de Asociaciones de Psicólogos (EFPA) para la evaluación de los tests. El modelo trata de poner a disposición de los usuarios información contrastada sobre las características teóricas, prácticas y psicométricas de los tests, facilitando con ello un mejor uso de las pruebas. **Método:** para llevar a cabo la revisión del modelo de evaluación de los tests se formó una comisión de trabajo en el seno de la EFPA formada por seis expertos de diferentes países que trabajaron en la actualización del modelo europeo previo, adaptándolo a los nuevos desarrollos en el ámbito de la evaluación psicológica y educativa. **Resultados:** la versión actualizada del modelo de la EFPA permite evaluar los tests de forma integral. En una primera parte se describe la prueba de forma exhaustiva, y en la segunda se lleva a cabo la evaluación cuantitativa y cualitativa de las características psicométricas de la prueba. **Conclusiones:** se presenta la revisión del modelo europeo para la descripción y evaluación de los tests y se comentan los resultados a la luz de los desarrollos recientes en el ámbito de la evaluación psicológica y educativa.

Palabras clave: revisión de tests, EFPA, calidad de los tests, uso de los tests.

Various national and international organisations such as the International Test Commission (ITC) and the European Federation of Psychologists' Association (EFPA) have been working for many years on diverse projects to improve the use of tests (Bartram, 2011; Evers et al., 2012; Muñiz & Bartram, 2007). Sensible test use requires the test to have pertinent psychometric properties, on the one hand, and, on the other, to be used adequately. As in any other scientific-technical area, the metric quality of a measuring instrument is a necessary condition to achieve rigorous results, but it is not enough, because inappropriate use of an instrument can ruin an excellent test. Aware of this fact, international organisations, such as the ITC and the EFPA, have developed a full set of projects aimed at improving the use of tests in applied settings. These projects are of a very diverse nature, although all of them can be structured around two large complementary strategies, one of a restrictive nature and the other one more informative.

The projects included in the restrictive strategy focus on limiting the use of tests to professionals who are trained and accredited for this activity (Bartram, 1996, 2011; Bartram & Coyne, 1998; Muñiz & Fernández-Hermida, 2000; Muñiz, Prieto, Almeida, & Bartram, 1999). In contrast, the informative strategy includes diverse types of projects aimed at disseminating information and knowledge about tests, assuming that the more information test users have, the more likely they are to use tests adequately. Within this strategy, various lines of action can be distinguished, among which are various sets of standards or guidelines, norms of the International Organization for Standardization (ISO), and test quality assessment.

Guidelines include all kinds of technical recommendations for the construction and analysis of tests (AERA, APA, NCME, 1999; Brennan, 2006; Downing & Haladyna, 2006; Wilson, 2005), or ethical and deontological standards (European Federation of Professional Psychologists' Associations, 2005; Fernández

Received: January 25, 2013 • Accepted: April 11, 2013
Corresponding author: Arne Evans
Dutch Association of Psychologists
Weesperplein 4
1018XA Amsterdam (Países Bajos)
e-mail: a.v.a.m.evers@uva.nl

Ballesteros et al., 2001; Joint Committee on Testing Practices, 2002; Koocher & Keith-Spiegel, 2007; Leach & Oakland, 2007; Lindsay, Koene, Ovreeide, & Lang, 2008), as well as other recommendations targeting specific settings, for example, for the translation and adaptation of tests (Hambleton, Merenda, & Spielberger, 2005), or others (Muñiz & Bartram, 2007). The ISO standards are not of a legal nature, but they establish a framework of international quality, which favours good professional practice. The new ISO 10667 standard, which regulates all the aspects involving assessment of people in work settings, was recently published (ISO, 2011). It includes the entire assessment process, constituting an excellent framework for the integration of other standards and guidelines such as those developed by the EFPA and ITC. Lastly, the third line of research focuses on the assessment of test quality, with the aim of providing users with all the necessary information for appropriate decision-making about measurement instruments. This is the framework of the model of test assessment presented in this work. Professionals of psychology have always demanded more technical information about tests (Evers et al., 2012; Muñiz & Bartram, 2007; Muñiz et al., 2001). This information can be provided in various ways, but to do this more systematically, the EFPA Committee on Tests and Testing developed the EFPA Review Model to assess the quality of tests (note that since 2011 the name of this committee has been changed to the EFPA Board of Assessment).

The current test review model was first published on the EFPA website in 2002 (Bartram, 2002), followed by several revisions of the model (e.g., Lindley, Bartram, & Kennedy, 2008). However, since the first publication of the model enormous advances have been made in the field of psychological and educational assessment, so it was necessary to revise the EFPA model thoroughly in order to include aspects of development derived from new information technologies in the area of assessment, such as computer-based tests, assessment by Internet, automated reports, or new models of Item Response Theory. The objective of this paper is to describe the revision of the model, which was completed in 2013.

Model development

The main goal of the EFPA Test Review Model is to provide a description and a detailed and rigorous assessment of the tests, scales and questionnaires used in the field of psychological and educational assessment. This information will be made available to test users and professionals, in order to improve tests and testing, and help them to make the right assessment decisions. The EFPA Test Review Model is part of the information strategy of the EFPA, which aims to provide all necessary technical information about the tests in order to enhance its use (Evers et al., 2012; Muñiz & Bartram, 2007). At present the EFPA model is in use in four European countries (Norway, Spain, Sweden, and the United Kingdom), whereas the model is translated, but not in use yet, in four other countries (Czech Republic, Hungary, Lithuania, and Russia) (Evers, 2012).

The original version of the EFPA test review model was produced from a number of sources, including the British Psychological Society (BPS) Test Review Evaluation Form (e.g., Bartram, Lindley, & Foster, 1990; Lindley et al., 2001), the Dutch Rating System for Test Quality (Evers, 2001a, 2001b), and the Spanish Questionnaire for the Evaluation of Psychometric Tests (Prieto & Muñiz, 2000). The synthesis of parts of these previously existing models was done with the permission of the associations concerned. Some major updated passages in the current revision have been adopted from the revised Dutch rating system (Evers, Lucassen, Meijer, & Sijtsma, 2010; Evers, Sijtsma, Lucassen, & Meijer, 2010) with permission of the authors. The revision also incorporates the guidance on reviewing translated and adapted test developed by Lindley (2009).

The revision of the model (version 4.2.6) was prepared by a Task Force of the EFPA Board of Assessment, whose members were Arne Evers (chair, the Netherlands), Carmen Hagemeister (Germany), Andreas Høstmælingen (Norway), Patricia Lindley (UK), José Muñiz (Spain), and Anders Sjöberg (Sweden). Also the convenor of the Board of Assessment, Dave Bartram, made a major contribution to the work of the Task Force. All (24) members of the Board of Assessment were consulted several times to comment on draft versions of the revised model. In addition, they were asked to circulate the draft revisions for review among experts in the countries they represent. All comments were discussed in the Task Force; some of them caused substantial changes, most of them were at least partially implemented.

Presentation of the model

In the Introduction to the EFPA Test Review Model, a test is defined as "any evaluative device or procedure in which a sample of examinee's behaviour in a specified domain is obtained and subsequently evaluated and scored using a standardized process" (AERA, APA, & NCME, 1999, p. 3). This definition is presented in order to make it clear that the review model applies to all such instruments, whether called a (single) scale, a (multi-scale) questionnaire, a projective technique, or whatever.

The EFPA Test Review Model is in three main parts; in the first part (Description of the instrument), all the features of the test are described in detail in a non-evaluative way. In the second part (Evaluation of the instrument), the fundamental properties of the test are evaluated against the EFPA model criteria (in Table 1 an overview of part 1 and 2 is presented). In the third part (Bibliography), the references used in the review model are included.

Description of the instrument

This part includes five sections: General description, Classification, Measurement and scoring, Computer generated reports, and Supply conditions and costs. The *General description*

Table 1
Outline of the EFPA Test Review Model Part 1 and 2

Part 1: Description of the instrument	**Part 2: Evaluation of the Instrument**
General description	Quality of the explanation of the rationale, the presentation and the information involved
Classification	Quality of the test materials
Measurement and scoring	Norms
Computer generated reports	Reliability
Supply conditions and costs	Validity
	Quality of computer generated reports
	Final evaluation

provides the basic information needed to identify the instrument and where to obtain it. It gives the title of the instrument, the publisher and/or distributor, the author(s), the date of original publication and the date of the version that is being reviewed. The second section dedicated to the *Classification* includes all the information needed in order to clearly classify the test, including aspects such as: Content domains, main areas of use, populations for which the test is intended, variables measured by the instrument, response mode, demands on the test taker, items format, ipsativity, number of items, administration mode, time required for administering the instrument, and availability of different forms. The third section on *Measurement and scoring* includes the scoring procedure for the test, a brief description of the scoring system to obtain global and partial scores, the scales used, and score transformation for norming scores. In the fourth section, *Computer generated reports*, the description of the characteristics of computerized reports include aspects such as media used, complexity, structure, sensitivity to context, clinical-actuarial, modifiability, transparency, style and tone, and intended recipients. Finally, in the fifth section, the *Supply conditions and costs* are described, specifying what the publisher will provide, to whom, under what conditions and at what costs. It defines the conditions imposed by the supplier on who may or may not obtain the instrument materials.

Evaluation of the instrument

This second part makes up the central part of the model, as all the essential aspects (such as norms or reliability, see Table 1) of the test are evaluated here. For each attribute, a number of items, sometimes arranged in sub-sections, has to be considered. Likert-type items make use of a scale with five categories ranging from zero to four, with the following meanings. Zero means that the attribute being evaluated cannot be rated as no, or insufficient information is provided, (1) means 'inadequate', (2) 'adequate', (3) 'good', and (4) 'excellent'. For each attribute, the reviewer is asked to comment upon the quality and also to integrate the ratings of the various items in an overall judgment of the adequacy of the attribute concerned. For these overall judgments, the same Likert-type scales are used. For these overall judgments, where a [0] or a [1] rating is provided on an attribute that is regarded as critical to the safe use of an instrument, the review will recommend that the instrument should not be used, except in exceptional circumstances by highly skilled experts or in research.

In order to carry out the evaluation, the more relevant sources of information are: (a) the manual and /or reports that are supplied by the publisher for the user, (b) open information that is available in the academic or other literature, (c) reports held by the publisher that are not formally published or distributed, and (d) reports that are commercial in confidence, as in some instances, publishers may have technically important material that they are unwilling to make public for commercial reasons. In the case of (d), publishers can be invited to provide this information under non-disclosure agreements for the information of the reviewers only.

Seven main attributes of the test are evaluated.

Quality of the explanation of the rationale, the presentation and the information provided

In this section, a number of ratings need to be given to various attributes of the documentation supplied with the instrument (or package). The term 'documentation' is taken to cover all those materials supplied or readily available to the qualified user: that is, the administrator's manual; technical handbooks; booklets of norms; manual supplements; updates from publishers/suppliers and so on.

Quality of the test materials

The quality of the materials of paper-and-pencil tests, computer based tests (CBT) and web based tests (WBT) are reviewed here. In the case of paper-and-pencil tests, aspects such as general quality of test materials (test booklets, answer sheets, test objects, etc.), ease with which the test taker can understand the task, or clarity and comprehensiveness of the instruction for the test taker are assessed in this section. In relation to CBT and WBT, special attention is paid to the software and usability of the interfaces.

Norms

This section is divided in two sub-sections, the first relative to norm-referenced tests, and the second to criterion-referenced tests. Special attention is paid in both parts to the nature, selection and characteristics of the sample(s) used. Also, the selection of critical scores and continuous norming procedures are analyzed (Bechger, Hemker, & Maris, 2009).

Reliability

Every aspect of test reliability is assessed in this section, with special emphasis on internal consistency, the test-retest coefficient, parallel forms, the item response theory approach, and inter-rater reliability. The size of reliability coefficients and sample adequacy are rated as well.

Validity

In order to properly understand the way the EFPA review model faces the assessment of validity, a few words have to be said. In the literature, many types of validity are differentiated, for instance, Drenth and Sijtsma (2006) mention eight different types. The differentiations may have to do with the purpose of validation or with the process of validation by specific techniques of data analysis. In the last decades of the past century, there was a growing consensus that validity should be considered as a unitary concept and that differentiations in types of validity should be considered as different types of gathering evidence only (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Borsboom, Mellenbergh, and Van Heerden (2004) state that a test is valid for measuring an attribute if variations in the attribute causally produce variation in the measured outcomes. Whichever approach to validity one prefers, for a standardised evaluation, it is necessary to have some structure for the concept of validity. For this reason, separate sub-sections on construct and criterion validity are included. Depending on the purpose of the test, one of these types of evidence may be more relevant than the other. However, it is considered that construct validity is the more fundamental concept and that evidence on criterion validity may add to establishing the construct validity of a test. It is also considered that a test may have different validities depending on the type of decisions made

with the test, the type of samples used, etc. However, inherent in a test review system is that a quality judgment has to be made about *the* (construct or criterion) validity of a test. This judgment should be a reflection of the quality of the evidence that the test can be used for the range of inferences and interpretations that are stated in the manual. The broader the intended applications, the more validity evidence the author/publisher should deliver. Note that the final ratings for construct and criterion validity will be a kind of average of this evidence and that there may be situations or groups for which the test may have higher or lower validities, or for which the validity may not have been studied at all. With this in mind, an exhaustive assessment of all types of validity evidences is carried out.

Quality of computer generated reports

Computer generated reports can be seen as varying in both their breadth and their specificity. Reports may also vary in the range of people for whom they are suitable. In some cases it may be that separate tailored reports are provided for different groups of recipients. In this section the most relevant features of the reports are evaluated, such as reliability, validity, fairness, acceptability, length, and overall adequacy. The evaluation can consider additional matters such as whether the reports take into account any checks of consistency of responding, response bias measures (e.g., measures of central tendency in ratings) and other indicators of the confidence with which the person's scores can be interpreted.

Final evaluation

This section contains a concise, clearly argued judgment about the test. It describes its pros and cons, and gives some general recommendations about how and when it might be used—together with warnings (where necessary) about when it should not be used. Depending on the evaluation of the critical technical criteria (norms, reliability, validity and computer generated reports), a recommendation is given for the type of use of the instrument and the required qualifications of the test user.

Main changes introduced by the revision of the EFPA model

In general, the changes concern four main issues. First, in the original version of the EFPA Review Model the focus of the more practical criteria (Quality of the explanation of the rationale, the presentation and the information provided, and Quality of the test materials) was primarily on paper-and-pencil tests (apart from the separate criterion for the evaluation of the quality of computer generated reports), but nowadays many new tests do not even have a paper-and-pencil version anymore. Second, in the original version of the model, the questions and recommendations regarding the psychometric criteria (norms, reliability and validity) were essentially based on classical test theory. However, in the last decade, more and more tests are constructed using non-classical approaches. Third, the model did not deal systematically with the effects of the continuing globalization on the world of tests and testing, such as the translation and adaption of tests for use in other cultures, testing of clients in a non-native language, the use of international norms, etc. Separate guidance on dealing with this was prepared for use with the original EFPA model (Lindley, 2009), but this needed to be formally incorporated into the model. Fourth, the model was a bit too focused on tests that are primarily used in work- and organizational settings. This is a reflection of the fact that the model had been primarily used for the review of tests used in these settings only. Although the scope of the model has become broader, for instance, the examples described were still almost exclusively work-centred.

The text of the original version of the model was screened by the members of the task force with respect to these four issues. If necessary, text was modified or content was added. In addition, many smaller changes (e.g., in the lay-out, some restructuring of the model, etc.) were implemented. The major changes with respect to the descriptive and evaluative sections are summarized in Table 2. This is intended to provide sufficient information to explain the changes in the descriptive section and for the evaluative criteria Quality of the explanation of the rationale, the presentation and the information provided, Quality of the test materials, and Quality of computer generated reports. The changes in the criteria Norms, Reliability, Construct validity, and Criterion validity require some further explanation.

Norms

Scoring a test usually results in a raw score. Raw scores partly are determined by characteristics of the test, such as number of items, time limits, item difficulty or item popularity (i.e., item mean score on positively phrased personality or attitude rating scales), and test conditions. Thus, the raw score is difficult to interpret and generally unsuited for practical use. To give meaning to a raw score two ways of scaling or categorizing raw scores can be distinguished (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). First, a set of scaled scores or norms may be derived from the distribution of raw scores of a reference group. This is called norm-referenced interpretation. Second, standards may be derived from a domain of skills or subject matter to be mastered (domain-referenced interpretation) or cut scores may be derived from the results of empirical validity research (criterion-referenced interpretation). With the latter two possibilities, raw scores will be categorized in two or more different score ranges, for example, to assign patients in different score ranges to different treatment programs, to assign pupils scoring below a critical score to remedial teaching, or to accept or reject applicants in personnel selection. The provision of norms, standards or cut scores is a basic requirement for the practical use of most tests but there are exceptions. Examples are tests for which only intra-individual interpretation is recommended (such as some ipsative scaled instruments) and tests used for intra-individual comparisons across time. In such cases the qualification 'not applicable' is used.

In the revised model, the section on norm-referenced interpretation has nine items, including four new items. Two of these new items are also introduced in the sub-sections on domain- and criterion-referenced norming. One item checks whether information on practice effects is available. This item will have no direct consequences for the overall judgment about the norms, but it is considered positively if some guidelines (or even norms in case of a retest) for the interpretation of scores are given, particularly for tests for which retesting is regular practice. The other item concerns the ageing of the norms. This item may have more severe consequences, as norms (or standards or cut scores) that are older than 20 years are considered 'inadequate'. Norms

Table 2
Summary of main changes in the EFPA Test Review Model

PART 1 DESCRIPTION OF THE INSTRUMENT	
Section 3: Classification	
Item 3.1	The number of content domains is extended and reordered.
Item 3.7	New item. This item asks for the capabilities and skills (e.g., vision, command of test language, reading) which are necessary for the test/taker to work on the test as intended and to allow for a fair interpretation.
Item 3.9	New item. If items with multiple choice mixed scale alternatives are used, this item checks whether the resulting scores are ipsative
Section 6: Supply conditions and costs	
Items 6.6 and 6.7	In these items, that relate to test-related and professional qualifications required by the supplier of the test, the EFPA competence levels are introduced.
PART 2 EVALUATION OF THE INSTRUMENT	
	Rating scale adjusted from 0 – 5 into 0 – 4 (the score of 2 was not used and left out).
Section 7: Quality of the explanation of the rationale, the presentation and the information provided	
Items 7.2.2.1 and 7.2.2.2	Item 7.2.2.1 deals with the adequacy of the description of the developmental process of a test. A second item is added, which asks for specific details if the test is developed through translation and/or adaptation.
Items 7.2.6 and 7.2.7	The one item with respect to the information regarding validity is split into two items, one for construct and one for criterion validity.
Item 7.2.8	New item. This item deals with the completeness of the information concerning computer generated reports.
Items 7.3.2 and 7.3.3	The one item with respect to the quality of the procedural instructions concerning scoring and norming was split into two items.
Item 7.3.7	New item. This item checks if restrictions on use (e.g. types of disability, literacy levels required) are mentioned in the manual.
Item 7.3.8	New item. This item asks for the completeness of the information concerning software and technical support in case of CBT or WBT.
Section 8: Quality of test materials	
This section is split in two sub-sections, one for paper-and-pencil tests and one for CBT or WBT. The two sub-sections have some items in common, but in the CBT/WBT-section also the quality of the software, the design of the user interface and the security of the test has to be judged.	
Old item 2.5	This item asked for the test quality of the local adaptation. It is deleted, because always the local adaptation is subject of a review.
Items 8.1.3 and 8.2.3	New items. These items ask for the clarity and comprehensiveness of the instruction for the test taker.
Sections 9, 10 & 11: Norms, Reliability and Validity	
The general introduction to these three sections is replaced by separate introductions for each chapter.	
The item that asks for an overall judgment of the technical information (i.e., norms, reliability ánd validity) is removed.	
Section 9: Norms	
Item 9.1.4	New item. In this item guidelines are given in case a model for continuous norming is used.
Item 9.1.5	An exhaustive list of procedures that can be used in sample selection is provided.
Item 9.1.6	New item. This item asks for the representativeness of the norm sample(s).
Item 9.1.8	New item. In this item the ageing/recency of the norms is judged.
Item 9.1.9	New item. This item asks if information about practice effects is supplied.
Sub-section 9.2	New sub-section. This sub-section deals with the quality of domain- or criterion-referenced norming.
Section 10: Reliability	
Item 10.2.2	New item. In this item the type of internal consistency coefficient that is used has to be marked.
Items 10.2.4, 10.3.4 and 10.4.4	New items. In these items is asked whether the samples used for computing the reliability coefficients concerned match the intended test takers.
Item 10.3.3	New item. In this item the test-retest interval has to be reported.
Item 10.4.2	New item. With respect to equivalence reliability it is judged to what extent assumptions for parallelism are met.
Sub-sections 10.5, 10.6 and 10.7	New sub-sections. Items with respect to reliability coefficients based on IRT, inter-rater reliability and other methods of reliability estimation are included.
Section 11.1: Construct validity	
Item 11.1.1	The list of research designs which can be used to study construct validity is extended.
Items 11.1.2 to 11.1.9	New items. For each design a separate item to judge whether the results support the construct validity is included.
Item 11.1.12	New item. This item asks for the age of the research.
Section 11.2: Criterion validity	
Item 11.2.3	New item. This item asks for the quality of the criterion measures.
Item 11.2.4	ROC-curves and guidelines for judging the outcomes of this type of assessing predictor-criterion relationships are introduced.
Item 11.2.5	New item. This item asks for the age of the research.
Section 12: Quality of computer generated reports	
Old item 2.12.5	The item dealing with the practicality of computer generated reports is deleted, because it was unclear and overlapped with item 12.6.
Item 12.6	The Length Index for judging the length of computer generated reports is removed.
Section 13: Final evaluation	
Recommendations	The rating for Validity-overall is used instead of the separate ratings for construct and criterion validity for the recommendation of the user level.
Recommendations	The number of recommendation categories is reduced from 7 to 6; EFPA User Qualification Levels are introduced.
Notes references and bibliography	The obligation for the reviewer to look for additional references about the test is removed.
Appendix	An aide memoire of critical points for comment when an instrument has been translated and/or adapted from a non-local context has been added.

that are younger than 10 years are considered 'excellent'. An item on the representativeness of the norm group(s) is now included, as this was not explicitly assessed with the former version of the model. Representativeness is primarily considered for the intended application domain. It is considered 'excellent' if data are gathered by means of a random sampling model and a thorough description of the composition of the sample(s) and the population(s) with respect to relevant background variables (such as gender, age, education, cultural background, occupation) is provided, whereas also good representativeness with regard to these variables is established.

The fourth new item sets guidelines for the sample size if a continuous norming approach is used. Opposite to the 'classical norming' approach, the continuous-norming procedure uses the information from all available groups to construct the norms for a specific group, which results in more accurate norms than classical norms (e.g., Zachary & Gorsuch, 1985). Thus, a continuous-norming procedure produces the same accuracy using smaller individual norm groups. Bechger, Hemker, and Maris (2009) studied a continuous-norming approach for eight groups and developed rules for the size of individual norm groups to be used in continuous norming. They used a linear regression approach assuming equal variances and standard-normal score distributions in all groups. To compare the accuracy of both approaches, they used the standard error of the mean. The results showed that a group size of about 70 in the continuous approach (for eight groups) produced the same accuracy as a group size of 200 in the classical approach, and that group sizes of 100 and 150 corresponded to sizes of 300 and 400, respectively. These group sizes are mean values but in the outer groups, accuracy is a bit worse than in the middle groups; hence, the outer groups should be larger whereas the middle groups may be smaller. However, the computed values are meant as an example only, as the required group sizes will differ if the number of groups differs and if a different continuous-norming approach is used (although linear regression is the most common approach). Therefore, test constructors are advised to supply evidence about the level of accuracy of the continuous norms for their test. They should also supply information on the other moments of the score distribution, as well as information about deviations from the statistical assumptions underlying their continuous-norming procedure

The new sub-section on domain-referenced interpretation has seven items. These items assess the selection and training of the experts, the number of experts used, the type of standard setting procedure, the type of coefficient for determining inter-rater agreement, the size of this coefficient, the ageing/recency of the norm study, and whether information about practice effects is available (the last two items being identical to those in the sections norm- and criterion-referenced interpretation).

In criterion-referenced interpretation, cut scores or expectancy tables are derived from empirical research. Actually, this concerns research on the criterion validity of the test, which also serves the process of setting norms empirically. Examples are research on the predictive validity of a test in personnel selection, and research on the sensitivity and specificity of a test in clinical psychology. This type of research is evaluated in the new sub-section on criterion-referenced norming exclusively from the perspective of setting norms; criterion validity per se is evaluated in the section concerned. The sub-section on criterion-referenced interpretation has three items. The first assesses the quality of the research. For the judgment of this item, no explicit guidelines are given as there is too much variation in the design of studies computing cut scores or expectancy tables. The other two items are the common ones for all three sub-sections (i.e., assessing ageing and the availability of information on practice effects).

Reliability

Reliability is a basic requirement for a test. However, different estimation methods may produce different reliability estimates and in different samples the test score may have different reliabilities, as reliability estimates depend on group heterogeneity. Thus, reliability results should be evaluated from the perspective of the test's application. Classical test theory assumes that a test score additively consists of a reliable component (also called true score) and a component caused by random measurement error. The objective of the reliability analysis is to estimate the degree to which test-score variance is due to true-score variance. In the previous version of the review model, recommendations were given with respect to internal consistency, test-retest reliability, and equivalence (parallel-form) reliability. In the revised model, coefficients based on IRT, inter-rater reliability, and other methods (e.g., generalizability theory, structural equation models) are also mentioned. As reliability estimates can differ depending on the characteristics of the group studied (particularly influential is the test-score variance), for some methods a question is added to the criteria that checks whether the samples used for computing the reliability coefficients match the intended test takers. In addition, if internal consistency coefficients are used the type of coefficient (e.g., alpha, lambda, *glb*) has to be marked; if test-retest reliability is reported, the length of the test-retest interval has to be mentioned; and if equivalence reliability is used, the extent to which assumptions for parallelism are met has to be judged. These additions stress the fact that variations in the design can also influence the outcomes of a reliability study.

For determining the overall rating for reliability, the reviewers are advised to take into account: (a) the nature of the instrument (e.g,. for some instruments internal consistency may be inappropriate, such as broad traits or scale aggregates), (b) the kind of decision based on the test score (e.g. for high-stakes decisions higher reliability coefficients are required than for low-stakes decisions), (c) whether one or more (types of) reliability studies are reported, (d) whether also standard errors of measurement are provided, (e) procedural issues (e.g. group size, number of reliability studies, heterogeneity of the group(s) on which the coefficient are computed), and (f) the comprehensiveness of the reporting on the reliability studies.

Construct validity

As argued above, in the revised version of the review model, the distinction between construct validity and criterion validity as separate criteria is maintained. Types of validity evidence that are construct-related (see below) are required for almost all tests, whatever the purpose of the test use (even when the purpose of a test is mere prediction, it would be odd not wanting to know what the test actually measures). Types of validity evidence that are criterion-related (see next section) will not be required or are less important for tests that are not intended for prediction.

Construct-related evidence should support the claim that the test measures the intended trait or ability. This concerns answers

to questions such as "What does the test measure?" and "Does the test measure the intended concept or does it partly or mainly measure something else?" By means of a great diversity of research designs, evidence for construct validity can be gathered. In the first item in this section, the reviewer has to mark which of the nine most common designs are used (a category 'other' is also included). These designs are: Exploratory Factor Analysis, Confirmatory Factor Analysis, (corrected) item-test correlations, testing for invariance of structure and differential item functioning across groups, differences between groups, correlations with other instruments and performance criteria (this may be research on criterion validity that is also relevant for construct validity), multi-trait-multi-method correlations, IRT methodology, or (quasi-) experimental designs. For each type of research, a separate question is formulated that asks for the adequacy of the results. In addition, but similar to the original version of the model, the sample sizes and the quality of the instruments used as criteria or markers have to be rated. Added to the model is an item in which the age of the research has to be filled in. Because ageing of research in one area may go faster than in another (depending on theoretical developments in that particular area), no general rule is formulated for taking the age of the research into account. It is left to the expertise of the reviewer to incorporate this information, like the methodological adequacy of the research, in his overall judgment about construct validity.

Criterion validity

Research on criterion-related evidence should demonstrate that a test score is a good predictor of non-test behavior or outcome criteria. Prediction can focus on the past (post-dictive or retrospective validity), the same moment in time (concurrent validity), or on the future (predictive validity). Basically, evidence of criterion validity is required for all kinds of tests. However, when it is explicitly stated in the manual that test use does not serve prediction purposes (such as educational tests that measure progress), criterion validity can be considered 'not applicable'.

The core item in this sub-section asks for the strength of the relation(s) that is found between the test and the criteria. In the original version of the model, only rules for rating the size of correlation coefficients were given. However, particularly for use in clinical situations, data on the sensitivity and the specificity of a test may give more useful information on the relation between a test and a criterion. ROC-curves are a popular way of quantifying the sensitivity and specificity. Therefore, recommendations for judging the outcomes of this way of investigating predictor-criterion relationships are added. Similar to the original version of the model, the size of the sample(s) used has to be rated, but an item has been added with which the quality of the criterion measure(s) is assessed. The quality of the criterion measure is dependent both on the reliability of the measure and the extent to which the measure represents the criterion construct. Similar to construct validity, an item which questions the age of the research is added. As for construct validity, it is left to the expertise of the reviewer to incorporate this information in his overall judgment.

Applying the model

As important as the model itself is the way in which it is applied. For the original version of the model, it was recommended that tests should be evaluated by two independent reviewers, in a peer review process similar to the usual evaluation of scientific papers and projects. A guide editor should oversee the reviews and may call in a third reviewer if significant discrepancies between the two reviews are found. Some variations in the procedure are possible, whilst ensuring the competence and independence of the reviewers, as well as the guide editor. EFPA recommends that the evaluations in these reviews are directed towards qualified practising test users, though they should also be of interest to academics, test authors and specialists in psychometrics and psychological testing. No changes are made to these procedural recommendations.

Another key issue is the publication of the results of the test evaluation. The basic idea is that the results are available for all professionals and users (either paid or for free). A good option is that reviews are made available on the website of the National Psychological Association, although they could also be published by third parties or in other media such as journals or books.

The intention of making this review model widely available is to encourage the harmonisation of review procedures and criteria across Europe. Although harmonisation is one of the objectives of the model, another objective is to offer a system for test reviews to countries in which test reviews are not common practice yet. It is realized that local issues may necessitate changes in the EFPA Test Review Model or in the procedures when countries start to use the Model. Therefore, the Model is called a *Model* to stress that local adaptations are possible to guarantee a better fit with local needs. Comments on the EFPA test review model are welcomed in the hope that the experiences of instrument reviewers will be instrumental in improving and clarifying the processes.

Future perspectives

The revision of the EFPA model for the description and evaluation of tests arises from the need to adapt the model to the advances undergone by measurement instruments in the field of psychological and educational assessment. New updates will certainly be necessary after some time because, fortunately, this field is continually progressing. We cannot see into the future—nobody can—but we comment below on some of the possible pathways for the future development of psychological and educational assessment, following the lines presented in Evers et al. (2012), Muñiz (2012) and Muñiz, Elosua and Hambleton (2013), and essentially focusing on tests. The great forces currently shaping psychological assessment are new information technologies, especially the advances in computer science, multimedia, and the Internet. Authors like Bennet (1999, 2006), Breithaupt, Mills and Melican (2006) or Drasgow, Luecht and Bennet (2006) think that new technologies are having special impact on all aspects involved in psychological assessment, such as test design, item construction, item presentation, test scoring, and tele-assessment. All of this is changing the format and content of assessment, and there are reasonable misgivings about whether paper-and-pencil tests, as we know them, will be capable of withstanding this new technological change. New ways of assessment emerge, but psychometric tests will continue to be essential tools, in view of their objectivity and economy in terms of means and time (Phelps, 2005, 2008).

According to Hambleton (2004, 2006), six large areas will attract the attention of researchers and professionals in the coming years. The *first* area is the international use of tests, due to increasing globalization and ease of communication, which

poses a whole series of problems concerning the adaptation of tests from one country to another (Byrne et al., 2009; Calvo et al., 2012; Hambleton et al., 2005; Nogueira et al., 2012; Ortiz et al., 2012). The *second* area is the use of new psychometric models and technologies to generate and analyze tests, especially the models of the Item Response Theory (Abad, Olea, Ponsoda, & García, 2011; De Ayala, 2009; Elosua, Hambleton, & Muñiz, 2013; Hambleton, Swaminathan, & Rogers, 1991; Muñiz, 1997; Van der Linden & Hambleton, 1997). The *third* area is the appearance of new item formats derived from important computer and multimedia advances (Irvine & Kyllonen, 2002; Shermis & Burstein, 2003; Sireci & Zenisky, 2006; Zenisky & Sireci, 2002). The *fourth* area that will claim a lot of attention is everything related to computerized tests and their relations with the Internet, with special mention of Computerised Adaptive Tests (Van der Linden & Glas, 2010; Olea, Abad, & Barrada, 2010). Remote assessment or tele-assessment is another line of research that is developing rapidly (Bartram & Hambleton, 2006; Leeson, 2006; Mills et al., 2002; Parshall et al., 2001; Wilson, 2005). The advances in automated trial scoring, which poses interesting challenges, are also noteworthy within this technological line (Shermis & Burstein, 2003; Williamson, Xi, & Breyer, 2012). The *fifth* area concerns the systems used to provide feedback (Goodman & Hambleton, 2004). Lastly, in the future, there will very probably be a great demand for *training* by diverse professionals who are related to assessment, not necessarily psychologists, but also psychologists, as well as teachers, doctors, nurses, etc.

These are some of the lines of research around which assessment activities will most probably revolve in the not too distant future, and they will have a great impact on the type of tests and their use. We did not intend to present an exhaustive review, but only to provide some clues to get one's bearings in the rapidly changing world of psychological assessment because, in future editions, the EFPA model for test assessment will have to echo the advances that occur.

Acknowledgements

The authors thank the other members of the EFPA Board of Assessment and the consultants in various European countries for their participation in the process of revision and for their valuable comments.

References

Abad, F.J., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en Ciencias Sociales y de la Salud* [Measurement in social health sciences]. Madrid: Síntesis.

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

Bartram, D. (1996). Test qualifications and test use in the UK: The competence approach. *European Journal of Psychological Assessment*, *12*, 62-71.

Bartram, D. (2002). *Review model for the description and evaluation of psychological tests*. Brussels: European Federation of Psychologists' Associations (EFPA).

Bartram, D. (2011). Contributions of the EFPA Standing Committee on Tests and Testing (SCTT) to standards and good practice. *European Psychologist, 16,* 149-159.

Bartram, D., & Coyne, I. (1998). Variations in national patterns of testing and test use: The ITC/EFPPA international survey. *European Journal of Psychological Assessment, 14*, 249-260.

Bartram, D., & Hambleton, R.K. (Eds.) (2006). *Computer-based testing and the Internet*. Chichester, UK: Wiley and Sons.

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

Bechger, T., Hemker, B., & Maris, G. (2009). *Over het gebruik van continue normering* [On the use of continuous norming]. Arnhem, The Netherlands: Cito.

Bennett, R.E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice*. *18*(3), 5-12.

Bennett, R.E. (2006). Inexorable and inevitable: The continuing story of technology and assessment. In D. Bartram & R.K. Hambleton (Eds.), *Computer-based testing and the Internet* (pp. 201-217). Chichester, UK: Wiley and Sons.

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061-1071.

Breithaupt, K.J., Mills, C.N., & Melican, G.J. (2006). Facing the opportunities of the future. In D. Bartram & R.K. Hambleton (Eds.), *Computer-based testing and the Internet* (pp. 219-251). Chichester, UK: Wiley and Sons.

Brennan, R.L. (Ed.) (2006). *Educational measurement*. Westport, CT: ACE/Praeger.

Byrne, B.M., Leong, F.T., Hambleton, R.K., Oakland, T., van de Vijver, F.J., & Cheung, F.M. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology, 3*(2), 94-105.

Calvo, N., Gutiérrez, F., Andión, O., Caseras, X., Torrubia, R., & Casas, M. (2012). Psychometric properties of the Spanish version of the self-report personality diagnostic questionnaire-4+ (PDQ-4+) in psychiatric outpatients. *Psicothema, 24*, 156-160.

De Ayala, R.J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Downing, S.M., & Haladyna, T.M. (Eds.) (2006). *Handbook of test development*. Hillsdale, NJ: Erlbaum.

Drasgow, F., Luecht, R.M., & Bennett, R.E. (2006). Technology and testing. In R.L. Brennan (Ed.), *Educational measurement* (pp. 471-515). Westport, CT: ACE/Praeger.

Drenth, P.J.D., & Sijtsma, K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen* (4e herziene druk) [Test theory. Introduction to the theory and application of psychological tests (4th revised ed.)]. Houten, The Netherlands: Bohn Stafleu van Loghum.

Elosua, P., Hambleton, R., & Muñiz, J. (2013). *Teoría de la respuesta al ítem aplicada con R* [Item response theory applied with R]. Madrid: La Muralla.

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

Evers, A. (2001a). Improving test quality in the Netherlands: Results of 18 years of test ratings. *International Journal of Testing*, *1*, 137-153.

Evers, A. (2001b). The revised Dutch rating system for test quality. *International Journal of Testing*, *1*, 155-182.

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

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de Kwaliteit van Tests (geheel herziene versie; gewijzigde herdruk)* [COTAN Rating system for test quality (completely revised edition; revised reprint)]. Amsterdam: NIP.

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J.R., et al. (2012). Testing practices in the 21st Century: Developments and European psychologists' opinions. *European Psychologist*, *17*, 300-319.

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

Fernández-Ballesteros, R., De Bruyn, E., Godoy, A., Hornke, L., Ter Laak, J., Vizcarro, C., et al. (2001). Guidelines for the assessment process (GAP): A proposal for discussion. *European Journal of Psychological Assessment, 17*, 187-200.

Goodman, D.P., & Hambleton, R.K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*, 145-220.

Hambleton, R.K. (2004). Theory, methods, and practices in testing for the 21st century. *Psicothema, 16*, 696-701.

Hambleton, R.K. (2006, March). *Testing practices in the 21st century*. Key Note Address, University of Oviedo, Spain.

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

Hambleton, R.K., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of item response theory*. Beverly Hills, CA: Sage.

Irvine, S., & Kyllonen, P. (Eds.) (2002). *Item generation for test development*. Mahwah, NJ: Erlbaum.

ISO (2011). *Procedures and methods to assess people in work and organizational settings (part 1 and 2)*. Geneva: Author.

Joint Committee on Testing Practices (2002). *Ethical principles of psychologists and code of conduct*. Washington, DC: Author.

Koocher, G., & Kith-Spiegel, P. (2007). *Ethics in psychology*. New York: Oxford University Press.

Leach, M., & Oakland, T. (2007). Ethics standards impacting test development and use: A review of 31 ethics codes impacting practices in 35 countries. *International Journal of Testing, 7*, 71-88.

Leeson, H.V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing, 6*, 1-24.

Lindley, P.A. (2009). *Reviewing translated and adapted tests*. Leicester, UK: British Psychological Society.

Lindley, P.A., Bartram, D., & Kennedy, N. (2008). *EFPA Review Model for the description and evaluation of psychological tests: Test review form and notes for reviewers: Version 3.42*. Brussels: EFPA Standing Committee on Tests and Testing (September, 2008).

Lindley, P.A. (Senior Editor), Cooper, J., Robertson, I., Smith, M., & Waters, S. (Consulting Editors) (2001). *Review of personality assessment instruments (Level B) for use in occupational settings. 2nd Edition*. Leicester, UK: BPS Books.

Lindsay, G., Koene, C., Ovreeide. H., & Lang, F. (Eds.) (2008). *Ethics for European psychologists*. Göttingen, Germany, and Cambridge, MA: Hogrefe.

Mills, C.N., Potenza, M.T., Fremer, J.J., & Ward, W.C. (Eds.) (2002). *Computer-based testing: Building the foundation for future assessments*. Hillsdale, NJ: Erlbaum.

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems* [Introduction to item response theory]. Madrid: Pirámide.

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

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

Muñiz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., et al. (2001). Testing practices in European countries. *European Journal of Psychological Assessment, 17*, 201-211.

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

Muñiz, J., & Fernández-Hermida, J.R. (2000). La utilización de los tests en España [Test use in Spain]. *Papeles del Psicólogo, 76*, 41-49.

Muñiz, J., Prieto, G., Almeida, L., & Bartram, D. (1999). Test use in Spain, Portugal and Latin American countries. *European Journal of Psychological Assessment, 15*, 151-157.

Nogueira, R., Godoy, A., Romero, P., Gavino, A., & Cobos, M.P. (2012). Propiedades psicométricas de la versión española del Obssesive Belief Questionnaire-Children Version (OBQ-CV) en una muestra no clínica [Psychometric properties of the Spanish version of the Obsessive Belief Questionnaire-Children's Version in a non-clinical sample]. *Psicothema, 24*, 674-679.

Olea, J., Abad, F., & Barrada, J.R. (2010). Tests informatizados y otros nuevos tipos de tests [Computerized tests and other new types of tests]. *Papeles del Psicólogo*, *31*, 94-107.

Ortiz, S., Navarro, C., García, E., Ramis, C., & Manassero, M.A. (2012). Validación de la versión española de la escala de trabajo emocional de Frankfurt [Validation of the Spanish version of the Frankfurt Emotion Work Scales]. *Psicothema, 24*, 337-342.

Parshall, C.G., Spray, J.A., Davey, T., & Kalohn, J. (2001). *Practical considerations in computer-based testing*. New York: Springer Verlag.

Phelps, R. (Ed.) (2005). *Defending standardized testing*. London: Erlbaum.

Phelps, R. (Ed.) (2008). *Correcting fallacies about educational and psychological testing*. Washington: American Psychological Association.

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

Shermis, M.D., & Burstein, J.C. (Eds.) (2003). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Erlbaum.

Sireci, S., & Zenisky, A.L. (2006). Innovative items format in computer-based testing: In pursuit of construct representation. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 329-348). Hillsdale, NJ: Erlbaum.

Van der Linden, W.J., & Glas, C.A.W. (Eds.) (2010). *Elements of adaptive testing*. London: Springer.

Van der Linden, W.J., & Hambleton, R.K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Williamson, D.M., Xi, X., & Breyer, J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2-13.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

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

Zenisky, A.L., & Sireci, S.G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*, 337-362.