EFPA REVIEW MODEL FOR THE DESCRIPTION AND EVALUATION OF PSYCHOLOGICAL AND EDUCATIONAL TESTS

TEST REVIEW FORM AND NOTES FOR REVIEWERS

VERSION 4.2.6

Version 4.2.6 is a major revision of Version 3-42 (2008) by a task force of the Board of Assessment of EFPA consisting of:

Arne Evers (chair, the Netherlands)
Carmen Hagemeister (Germany)
Andreas Høstmælingen (Norway)
Patricia Lindley (UK)
José Muñiz (Spain)
Anders Sjöberg (Sweden)

Approved by the EFPA General Assembly, 13-07-2013

CONTENTS

1 Introduction

The main goal of the EFPA Test Review Model is to provide a description and a detailed and rigorous assessment of the psychological assessment tests, scales and questionnaires used in the fields of Work, Education, Health and other contexts. 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 evaluations of all necessary technical information about tests in order to enhance their use (Evers et al., 2012; Muñiz & Bartram, 2007). Following the *Standards for Educational and Psychological Testing* the label *test* is used for 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" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 3). Therefore, this review model applies to all instruments that are covered under this definition, whether called a scale, questionnaire, projective technique, or whatever.

The original version of the EFPA test review model was produced from a number of sources, including the BPS Test Review Evaluation Form (developed by Newland Park Associates Limited, NPAL, and later adopted by the BPS Steering Committee on Test Standards); the Spanish Questionnaire for the Evaluation of Psychometric Tests (developed by the Spanish Psychological Association) and the Rating System for Test Quality (developed by the Dutch Committee on Tests and Testing of the Dutch Association of Psychologists). Much of the content was adapted with permission from the review proforma originally developed in 1989 by Newland Park Associates Ltd for a review of tests used by training agents in the UK (see Bartram, Lindley & Foster, 1990). This was subsequently used and further developed for a series of BPS reviews of instruments for use in occupational assessment (e.g., Bartram, Lindley, & Foster, 1992; Lindley et al., 2001). The first version of the EFPA review model was compiled and edited by Dave Bartram (Bartram, 2002a, 2002b) following an initial EFPA workshop in March 2000 and subsequent rounds of consultation. A major update and revision was carried out by Patricia Lindley, Dave Bartram, and Natalie Kennedy for use in the BPS review system (Lindley et al, 2004). This was subsequently adopted by EFPA in 2005 (Lindley et al., 2005) with minor revisions in 2008 (Lindley et al., 2008). The current version of the model has been prepared by a Task Force of the EFPA Board of Assessment, whose members are Arne Evers (Chair, the Netherlands), Carmen Hagemeister (Germany), Andreas Høstmælingen (Norway), Patricia Lindley (UK), José Muñiz (Spain), and Anders Sjöberg (Sweden). In this version the notes and checklist for translated and adapted tests produced by Pat Lindley and the Consultant Editors of the UK test reviews have been integrated (Lindley, 2009). The texts of some major updated passages are based on the revised Dutch rating system for test quality (Evers, Lucassen, Meijer, & Sijtsma, 2010; Evers, Sijtsma, Lucassen, & Meijer, 2010).).

The EFPA test review model is divided into three main parts. In the first part (Description of the instrument) all the features of the test evaluated are described in detail. In the second part (Evaluation of the instrument) the fundamental properties of the test are evaluated: Test materials, norms, reliability, validity, and computer generated reports, including a global final evaluation. In the third part (Bibliography), the references used in the review are included.

As important as the model itself is the proper implementation of the model. The current version of the model is intended for use by two independent reviewers, in a peer review process similar to the usual evaluation of scientific papers and projects. A consulting editor will 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 consulting 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.

Another key issue is the publication of the results of a test's evaluation. The results should be available for all professionals and users (either paid or for free). A good option is that results are 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 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 which do not have their own review procedures. It is realized that local issues may necessitate changes in the EFPA Test Review Model or in the review 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 users will be instrumental in improving and clarifying the processes.

PART 1 DESCRIPTION OF THE INSTRUMENT

2 General description

This section of the form should provide the basic information needed to identify the instrument and where to obtain it. It should give 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 questions 2.1.1 through 2.7.3 should be straightforward. They are factual information, although some judgment will be needed to complete information regarding content domains.

	Reviewer[1]	
	Date of current review	
	Date of previous review *(if applicable)*[2]	
2.1.1	**Instrument name (local version)**	
2.1.2	**Shortname of the test** *(if applicable)*	
2.2	**Original test name** *(if the local version is an adaptation)*	
2.3	**Authors of the original test**	
2.4	**Authors of the local adaptation**	
2.5	**Local test distributor/publisher**	
2.6	**Publisher of the original version of the test** *(if different to current distributor/publisher)*	
2.7.1	**Date of publication of current revision/edition**	
2.7.2	**Date of publication of adaptation for local use**	
2.7.3	**Date of publication of original test**	

[1] Each country can decide either to publish the reviewers' names when the integrated review is published or to opt for anonymous reviewing.
[2] This information should be filled in by the editor or the administration.

General description of the instrument Short stand-alone non-evaluative description (200-600 words)

A concise non-evaluative description of the instrument should be given here. The description should provide the reader with a clear idea of what the instrument claims to be - what it contains, the scales it purports to measure etc. It should be as neutral as possible in tone. It should describe what the instrument is, the scales it measures, its intended use, the availability and type of norm groups, general points of interest or unusual features and any relevant historical background. This description may be quite short (200-300 words). However, for some of the more complex multi-scale instruments, it will need to be longer (300-600 words). It should be written so that it can stand alone as a description of the instrument. As a consequence it may repeat some of the more specific information provided in response to sections 2 – 6. It should outline all versions of the instrument that are available and referred to on subsequent pages.

This item should be answered from information provided by the publisher and checked for accuracy by the reviewer.

3 Classification

3.1	**Content domains** *(select all that apply)* You should identify the content domains specified by the publisher. Where these are not clear, this should be indicated and you should judge from the information provided in the manual (standardisation samples, applications, validation etc.) what the most appropriate answers are for 3.1.	□ Ability - General □ Ability - Manual skills/dexterity □ Ability - Mechanical □ Ability Learning/memory □ Ability - Non-verbal/abstract/inductive □ Ability - Numerical □ Ability - Perceptual speed/checking □ Ability - Sensorimotor □ Ability Spatial/visual □ Ability - Verbal □ Attention/concentration □ Beliefs □ Cognitive styles □ Disorder and pathology □ Family function □ Group function □ Interests □ Motivation □ Organisational function, aggregated measures, climate etc □ Personality – Trait □ Personality – Type □ Personality – State □ Quality of life □ Scholastic achievement (educational test) □ School or educational function □ Situational judgment □ Stress/burnout □ Therapy outcome □ Values □ Well-being □ Other (please describe):
3.2	**Intended or main area(s) of use** *(please select those that apply)* You should identify the intended areas of uses specified by the publisher. Where these are not clear, this should be indicated and you should judge from the information provided in the manual (standardisation samples, applications, validation etc) what the most appropriate answers are for 3.2.	□ Clinical □ Advice, guidance and career choice □ Educational □ Forensic □ General health, life and well-being □ Neurological □ Sports and Leisure □ Work and Occupational □ Other (please describe):
3.3	**Description of the populations for which the test is intended** This item should be answered from information provided by the publisher. For some tests this may be very general (e.g. adults), for others it may be more specific (e.g. manual workers, or boys	

	aged 10 to 14). Only the stated populations should be mentioned here. Where these may seem inappropriate, this should be commented on in the Evaluation part of the review.	
3.4	**Number of scales and brief description of the variable(s) measured by the instrument** This item should be answered from information provided by the publisher. Please indicate the number of scales (if more than one) and provide a brief description of each scale if its meaning is not clear from its name. Reviews of the instrument should include discussion of other derived scores where these are commonly used with the instrument and are described in the standard documentation - e.g. primary trait scores as well as Big Five secondary trait scores for a multitrait personality test, or subtest, factor and total scores on an intelligence test.	
3.5	**Response mode** This item should be answered from information provided by the publisher. If any special pieces of equipment (other than those indicated in the list of options, e.g. digital recorder) are required, they should be described here. In addition, any special testing conditions should be described. 'Standard testing conditions' are assumed to be available for proctored/supervised assessment. These would include a quiet, well-lit and well-ventilated room with adequate desk-space and seating for the necessary administrator(s) and candidate(s).	□ Oral interview □ Paper & pencil □ Manual (physical) operations □ Direct observation □ Computerised □ Other (indicate):
3.6	**Demands on the test taker** This item should be answered from information provided by the publisher. Which capabilities and skills are necessary for the test taker to work on the test as intended and to allow for a fair interpretation of the test score? It is usually clear if a total lack of some prerequisite impairs the ability to complete the test (such as being blind and being given a normal paper-and-pencil test) but the requirements listed should be classified as follows: • "Irrelevant / not necessary" means that this skill is not necessary at all – such as manual capabilities to answer oral ques-	Manual capabilities (*select one*) □ irrelevant / not necessary □ necessary information given □ information missing Handedness (*select one*) □ irrelevant / not necessary □ necessary information given □ information missing Vision (*select one*) □ irrelevant / not necessary □ necessary information given □ information missing Hearing (*select one)*

	tions verbally. • "Necessary information given" means that the possible amount of limitation is stated. • "Information missing" means that there might be limitations on test users without the specific capability or skill (known from theory or empirical results) but this is not clear from information provided by the test publisher e.g. if the test uses language that is not the test taker's first language.	□ irrelevant / not necessary □ necessary information given □ information missing Command of test language (understanding and speaking) (*select one*) □ irrelevant / not necessary □ necessary information given □ information missing Reading (*select one*) □ irrelevant / not necessary □ necessary information given □ information missing Writing (*select one*) □ irrelevant / not necessary □ necessary information given □ information missing
3.7	**Items format** *(select one)* This item should be answered from information provided by the publisher. Two types of multiple choice formats are differentiated. The first type concerns tests in which the respondent has to select the right answer from a number of alternatives as in ability testing (e.g., a figural reasoning test). The second type deals with questionnaires in which there is no clear right answer. This format requires test takers to make choices between sets of two or more items drawn from different scales (e.g., scales in a vocational interest inventory or a personality questionnaire). This format is also called 'multidimensional', because the alternatives belong to different scales or dimensions. In this case it is possible that the statements have to be ranked or the most- and least-like-me options be selected. This format may result in ipsative scales (see question 3.8). In Likert scale ratings the test taker also has to choose from a number of alternatives, but the essential difference with the multiple choice format is that the scales used are unidimensional (e.g., ranging from 'never' to 'always' or from 'very unlikely' to 'very likely') and that the test taker does not have to choose between alternatives from different dimensions. A scale should also be marked as a Likert scale when there are only two alternatives on one dimension (e.g., yes/no or always/never).	□ Multiple choice (ability testing, or right/wrong) Number of alternatives: .... □ Multiple choice (mixed scale alternatives) Number of alternatives: …. □ Likert scale ratings Number of alternatives: …. □ Open □ Other (please describe)

3.8	**Ipsativity** As mentioned in 3.7 multiple choice mixed scale alternatives *may* result in ipsative scores. Distinctive for ipsative scores is that the score on each scale or dimension is constrained by the scores on the other scales or dimensions. In fully ipsative instruments the sum of the scale scores is constant for each person. Other scoring procedures can result in ipsativity (e.g. subtraction of each person's overall mean from each of their scale scores)	□ Yes, multiple choice mixed scale alternatives resulting in partially or fully ipsative scores □ Yes, other item formats with scoring procedures resulting in partially or fully ipsative scores □ No, multiple choice mixed scale alternatives NOT resulting in ipsative scores □ Not relevant
3.9	**Total number of test items and number of items per scale or subtest** This item should be answered from information provided by the publisher. If the instrument has several scales or subtests, indicate the total number of items and the number of items for each scale or subtest. Where items load on more than one scale or subtest, this should be documented.	
3.10	**Intended mode of use (conditions under which the instrument was developed and validated)** (*select all that apply*) This item is important as it identifies whether the instrument has been designed with the intention of it being used in unsupervised or uncontrolled administration conditions. Note that usage modes may vary across versions of a tool. This item should be answered from information provided by the publisher and checked for accuracy. Note. The four modes are defined in the *International Guidelines on Computer-Based and Internet Delivered Testing* (International Test Commission, 2005, pp. 5-6).	□ *Open mode*: Where there is no direct human supervision of the assessment session and hence there is no means of authenticating the identity of the test-taker. Internet-based tests without any requirement for registration can be considered an example of this mode of administration. □ *Controlled mode*: No direct human supervision of the assessment session is involved but the test is made available only to known test-takers. Internet tests will require test-takers to obtain a logon username and password. These often are designed to operate on a one-time-only basis. □ *Supervised (proctored) mode:* Where there is a level of direct human supervision over test-taking conditions. In this mode test-taker identity can be authenticated. For Internet testing this would require an administrator to log-in a candidate and confirm that the test had been properly administered and completed. □ *Managed mode:* Where there is a high level of human supervision and control over the test-taking environment. In CBT testing this is normally achieved by the use of dedicated testing centres, where there is a high level of control over access, security, the qualification of test administration staff and the quality and technical specifications of the test equipment.

3.11	**Administration mode(s)** *(select all that apply)* This item should be answered from information provided by the publisher. If any special pieces of equipment (other than those indicated in the list of options, e.g. digital recorder) are required, they should be described here. In addition, any special testing conditions should be described. 'Standard testing conditions' are assumed to be available for proctored/ supervised assessment. These would include a quiet, well-lit and well-ventilated room with adequate desk-space and seating for the necessary administrator(s) and candidate(s).	□ Interactive individual administration □ Supervised group administration □ Computerised locally-installed application – supervised/proctored □ Computerised web-based application – supervised/proctored □ Computerised locally-installed application – unsupervised/self-assessment □ Computerised web-based application – unsupervised/self-assessment □ Other (indicate):
3.12	**Time required for administering the instrument** *(please specify for each administration mode)* This item should be answered from information provided by the publisher. The response to this item can be broken down into a number of components. In most cases, it will only be possible to provide general estimates of these rather than precise figures. The aim is to give the potential user a good idea of the time investment associated with using this instrument. Do NOT include the time needed to become familiar with the instrument itself. Assume the user is experienced and qualified. • Preparation time (the time it takes the administrator to prepare and set out the materials for an assessment session; access and login time for an online administration). • Administration time per session: this includes the time taken to complete all the items and an estimate of the time required to give instructions, work through example items and deal with any debriefing comments at the end of the session. • Scoring: the time taken to obtain the raw-scores. In many cases this may be automated. • Analysis: the time taken to carry out further work on the raw scores to derive other measures and to produce a reasonably comprehensive interpretation (assuming you are familiar with the instrument). Again, this may be automated. • Feedback: the time required to prepare and provide feedback to a test taker and other stakeholders.	Preparation: Administration: Scoring: Analysis: Feedback:

	It is recognised that time for the last two components could vary enormously - depending on the context in which the instrument is being used. However, some indication or comments will be helpful.	
3.13	**Indicate whether different forms of the instrument are available and which form(s) is (are) subject of this review** Report whether or not there are alternative versions (genuine or pseudo-parallel forms, short versions, computerised versions, etc.) of the instrument available and describe the applicability of each form for different groups of people. In some cases, different forms of an instrument are meant to be equivalent to each other - i.e. alternative forms. In other cases, various forms may exist for quite different groups (e.g. a children's form and an adult's form). Where more than one form exists, indicate whether these are equivalent/alternate forms, or whether they are designed to serve different functions - e.g. short and long version; ipsative and normative version. Also describe whether or not parts of the whole test can be used instead of the whole instrument. If computerised versions do exist, describe briefly the software and hardware requirements. Note that standalone computer based tests (CBT) and online packages, if available, should be indicated.	

4 Measurement and scoring

4.1	**Scoring procedure for the test** *(select all that apply)* This item should be completed by reference to the publisher's information and the manuals and documentation. Bureau services are services provided by the supplier - or some agent of the supplier - for scoring and interpretation. In general these are optional services. If scoring and/or interpretation can be carried out ONLY through a bureau service, then this should be stated in the review - and the costs included in the recurrent costs item.	□ Computer scoring with direct entry of responses by test taker □ Computer scoring by Optical Mark Reader entry of responses from the paper response form □ Computer scoring with manual entry of responses from the paper response form □ Simple manual scoring key – clerical skills only required □ Complex manual scoring – requiring training in the scoring of the instrument □ Bureau-service – e.g. scoring by the company selling the instrument □ Other (please describe):
4.2	**Scores** This item should be completed by reference to the publisher's information and the manuals and documentation. Brief description of the scoring system to obtain global and partial scores, correction for guessing, qualitative interpretation aids, etc).	
4.3	**Scales used** (*select all that apply)* This item should be completed by reference to the publisher's information and the manuals and documentation.	*Percentile Based Scores* □ Centiles □ 5-grade classification: 10:20:40:20:10 centile splits □ Deciles □ Other (please describe): *Standard Scores* □ Z-scores □ IQ deviation quotients etc (e.g. mean 100, SD=15 for Wechsler or 16 for Stanford-Binet) □ College Entrance Examination Board (e.g. SAT mean=500, SD=100) □ Stens □ Stanines, C-scores □ T-scores □ Other (please describe): □ Critical scores, expectancy tables or other specific decision oriented indices □ Raw score use only □ Other (please describe):

4.4	**Score transformation for standard scores**	□ Normalised – standard scores obtained by use of normalisation look-up table □ Not-normalised – standard scores obtained by linear transformation □ Not applicable

5 Computer generated reports

Note that this section is purely *descriptive*. Evaluations of the reports should be given in the Evaluation part of the review

For instances where there are multiple generated reports available please complete items 5.2 – 5.13 for each report or substantive report section (copy pages as necessary). This classification system could be used to describe two reports provided by a system, for example, Report 1 may be intended for the test taker or other un-trained users, and Report 2 for a trained user who is competent in the use of the instrument and understands how to interpret it.

5.1	**Are computer generated reports available with the instrument?** If the answer to 5.1 is 'YES' then the following classification should be used to classify the types of reports available. For many instruments, there will be a range of reports available. Please complete a separate form for each report	□ Yes (complete items below) □ No (move to item 6.1)
5.2	**Name or description of report** *(see introduction to this section)*	
5.3	**Media** *(select all that apply)* Reports may consist wholly of text or contain text together with graphical or tabular representations of scores (e.g. sten profiles). Where both text and data are presented, these may simply be presented in parallel or may be linked, so that the relationship between text statements and scores is made explicit.	□ Text only □ Unrelated text and graphics □ Integrated text and graphics □ Graphics only
5.4	**Complexity** *(select one)* Some reports are very simple, for example just substituting a text unit for a sten score in a scale-by-scale description. Others are more complex, involving text units which relate to patterns or configurations of scale scores and which consider scale interaction effects.	□ Simple (For example, a list of paragraphs giving scale descriptions) □ Medium (A mixture of simple descriptions and some configural descriptions) □ Complex (Contains descriptions of patterns and configurations of scale scores, and scale interactions)
5.5	**Report structure** *(select one)* Structure is related to complexity.	□ Scale based – where the report is built around the individual scales. □ Factor based – where the report is constructed around higher order factors - such as the 'Big Five' for personality measures. □ Construct based – where the report is built around one or more sets of constructs (e.g. in a work setting these could be such as team types, leadership styles, or tolerance to stress; in a clinical setting these could be different kinds of psychopathology; etc.) which are

		linked to the original scale scores. □ Criterion based where the reports focuses on links with empirical outcomes (e.g. school performance, therapy outcome, job performance, absenteeism etc). □ Other (please describe):
5.6	**Sensitivity to context** *(select one)* When people write reports they tailor the language, form and content of the report to the person who will be reading it and take account of the purpose of the assessment and context in which it takes place. In a work and organizational context a report produced for selection purposes will be different from one written for guidance or development; a report for a middle-aged manager will differ from that written for a young person starting out on a training scheme and so on. In an educational context a report produced for evaluation of a students' global ability to learn and function in a learning environment will be different from a report produced to assess whether or not a student has a specific learning disorder. A report directed to other professionals suggesting learning goals and interventions will differ from reports directed to parents informing them of their child's strengths and weaknesses. In a clinical context a report produced for diagnostic purposes will be different from a report evaluating a patient's potential for risk-taking behaviour. A report produced with the purpose of providing feedback to patients will be different from a report produced with the purpose of informing authorities whether or not it is safe to release a patient from involuntary treatment.	□ One version for all contexts □ Pre-defined context-related versions; number of contexts: .... □ User definable contexts and editable reports
5.7	**Clinical-actuarial** *(select all that apply)* Most report systems are based on clinical judgment. That is, one or more people who are 'expert-users' of the instrument in question will have written the text units. The reports will, therefore, embody their particular interpretations of the scales. Some systems include actuarial reports where the statements are based on empirical validation studies linking scale scores to, for example, job performance measures, clinical classification, etc.	□ Based on clinical judgment of one expert □ Based on clinical judgment of group of experts □ Based on empirical/actuarial relationships
5.8	**Modifiability** (*select one)*	□ Not modifiable (fixed print-only output) □ Limited modification (limited to certain areas,

	The report output is often fixed. However, some systems will produce output in the form of a file that can be processed by the user. Others may provide online interactive access to both the end user and the test taker.	e.g. biodata fields) □ Unlimited modification (e.g. through access to Word processor document file) □ Interactive report which provides test taker with an opportunity to insert comments or provides ratings of accuracy of content (e.g. through shared online access to an interactive report engine)
5.9	**Degree of finish** *(select one)* Extent to which the system is designed to generate integrated text - in the form of a ready-to-use report - or a set of 'notes', comments, hypotheses etc..	□ Publication quality □ Draft quality
5.10	**Transparency** *(select one)* Systems differ in their openness or transparency to the user. An open system is one where the link between a scale score and the text is clear and unambiguous. Such openness is only possible if both text and scores are presented and the links between them made explicit. Other systems operate as 'black boxes', making it difficult for the user to relate scale scores to text.	□ Clear linkage between constructs, scores and text □ Concealed link between constructs, scores and text □ Mixture of clear/concealed linkage between constructs, scores and text
5.11	**Style and tone** *(select one)* Systems also differ in the extent to which they offer the report reader guidance or direction. In a work and organizational context a statement as "Mr X is very shy and will not make a good salesman..." is stipulative, whereas other statements are designed to suggest hypotheses or raise questions, such as "From his scores on scale Y, Mr X appears to be very shy compared to a reference group of salespersons. If this is the case, he could find it difficult working in a sales environment. This needs to be explored further with him". In an educational context a stipulative statement might be: "The results show that X's mathematical skills are two years below the average of his peers", whereas a statement designed to suggest hypotheses might be: "The results indicate X is easily distracted by external stimuli while performing tasks. Behavioural observations during testing support this. This should be taken under consideration when designing an optimal learning environment for X". In a clinical context a stipulative statement might be: "Test scores indicate the patient has severe visual neglect, and is not able to safely operate a motor vehicle", whereas a state-	□ Directive/stipulative □ Guidance/suggests hypotheses □ Other (please describe):

	ment designed to suggest hypotheses might be: "Mrs X's test scores indicate she may have problems establishing stable emotional relationships. This should be explored further before a conclusion regarding diagnosis is drawn".	
5.12	**Intended recipients** (*select all that apply)* Reports are generally designed to address the needs of one or more categories of users. Users can be divided into four main groups: *a) Qualified test users.* These are people who are sufficiently knowledgeable and skilled to be able to produce their own reports based on scale scores. They should be able to make use of reports that use technical psychometric terminology and make explicit linkages between scales and descriptions. They should also be able to customize and modify reports. *b) Qualified system users.* While not competent to generate their own reports from a set of scale scores, people in this group are competent to use the outputs generated by the system. The level of training required to attain this competence will vary considerably, depending on the nature of the computer reports (e.g. trait-based versus competency-based, simple or complex) and the uses to which its reports are to be put (low stakes or high stakes). *c) Test Takers.* The person who takes the instrument will generally have no prior knowledge of either the instrument or the type of report produced by the system. Reports for them will need to be in language that makes no assumptions about psychometric or instrument knowledge. *d) Third parties.* These include people - other than the candidate - who will be privy to the information presented in the report or who may receive a copy of the report. They may include potential employers, a person's manager or supervisor or the parent of a young person receiving careers advice. The level of language required for people in this category would be similar to that required for reports intended for Test Takers.	□ Qualified test users □ Qualified system users □ Test takers □ Third Parties
5.13	**Do distributors offer a service to modify and/or develop customised computerised reports?** *(select one)*	□ Yes □ No

6 Supply conditions and costs

This defines 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. If one of the options does not fit the supply conditions, provide a description of the relevant conditions

6.1	**Documentation provided by the distributor as part of the test package** *(select all that apply)*	□ User Manual □ Technical (psychometric) manual □ Supplementary technical information and updates (e.g. local norms, local validation studies etc.) □ Books and articles of related interest
6.2	**Methods of publication** *(select all that apply)* For example, technical manuals may be kept up-to-date and available for downloading from the Internet, while user manuals are provided in paper form or on a CD/DVD.	□ Paper □ CD or DVD □ Internet download □ Other (specify):
Items 6.3 - 6.5 cover costs. This information is likely to be the most quickly out of date. It is recommended that the supplier or publisher is contacted as near the time of publication of the review as possible, to provide current information for these items.		
6.3.1	**Start-up costs** Price of a complete set of materials (all manuals and other material sufficient for at least one sample administration). Specify how many test takers could be assessed with the materials obtained for start-up costs, and whether these costs include materials for recurrent assessment. This item should try to identify the 'set-up' cost. That is the costs involved in obtaining a full reference set of materials, scoring keys and so on. It only includes training costs if the instrument is a 'closed' one - where there will be an unavoidable specific training cost, regardless of the prior qualification level of the user. In such cases, the training element in the cost should be made explicit. The initial costs do NOT include costs of general-purpose equipment (such as computers, DVD players and so on). However, the need for these should be mentioned. In general, define: any special training costs; costs of administrator's manual; technical manual(s); specimen or reference set of materials; initial software costs, etc.	

6.3.2	**Recurrent costs** Specify, where appropriate, recurrent costs of administration and scoring separately from costs of interpretation (see 6.4.1 – 6.5). This item is concerned with the on-going cost of using the instrument. It should give the cost of the instrument materials (answer sheets, non-reusable or reusable question booklets, profile sheets, computer usage release codes or 'dongle' units, etc.) per person per administration. Note that in most cases, for paper-based administration such materials are not available singly but tend to be supplied in packs of 10, 25 or 50. Itemise any annual or per capita licence fees (including software release codes where relevant), costs of purchases or leasing re-usable materials, and per candidate costs of non-reusable materials.	
6.4.1	**Prices for reports generated by user installed software**	
6.4.2	**Prices for reports generated by postal/fax bureau service**	
6.4.3	**Prices for reports by Internet service**	
6.5	**Prices for other bureau services: correcting or developing automatic reports**	
6.6	**Test-related qualifications required by the supplier of the test** *(select all that apply)* This item concerns the user qualifications required by the supplier. For this item, where the publisher has provided user qualification information, this should be noted against the categories given. Where the qualification requirements are not clear this should be stated under 'Other' *not* under 'None'. 'None' means that there is an explicit statement regarding the lack of need for qualification. For details of the EFPA Level 2 standard, consult the latest version of these on the EFPA website.	□ None □ Test specific accreditation □ Accreditation in general achievement testing: measures of maximum performance in attainment (equivalent to EFPA Level 2) □ Accreditation in general ability and aptitude testing: measures of maximum performance in relation to potential for attainment (equivalent to EFPA Level 2) □ Accreditation in general personality and assessment: measures of typical behaviour, attitudes and preferences (equivalent to EFPA Level 2) □ Other (specify):

6.7	**Professional qualifications required for use of the instrument** *(select all that apply)* This item concerns the user qualifications required by the supplier. For this section, where the publisher has provided user qualification information, this should be noted against the categories given. Where the qualification requirements are not clear this should be stated under 'Other' *not* under 'None'. 'None' means that there is an explicit statement regarding the lack of need for qualification. For details of the EFPA user standards, consult the latest version of these on the EFPA website.	□ None □ Practitioner psychologist with qualification in the relevant area of application □ Practitioner psychologist □ Research psychologist □ Non-psychologist academic researcher □ Practitioner in relevant related professions (therapy, medicine, counselling, education, human resources etc.). Specify: …… □ EFPA Test User Qualification Level 1 or national equivalent □ EFPA Test User Qualification Level 2 or national equivalent □ Specialist qualification equivalent to EFPA Test User Standard Level 3 □ Other (indicate):

PART 2 EVALUATION OF THE INSTRUMENT

Sources of information

Potentially there are four sources of information that might be consulted in carrying out this evaluation:

1. The manual and /or reports that are supplied by the publisher for the user:
   These are always supplied by the publisher /distributor before the instrument is accepted by the reviewing organisation and form the core materials for the review.

2. Open information that is available in the academic or other literature:
   This is generally sourced by the reviewer and the reviewer may make use of this information in the review and the instrument may be evaluated as having (or having not) made reference to the information in its manual.

3. Information held by the publisher that is not formally published or distributed:
   The distributor/publisher may make this available at the outset or may send it when the review is sent back to the publisher to check for factual accuracy. The reviewer should make use of this information but note very clearly at the beginning of the comments on the technical information that "the starred rating in this review refers to materials held by the publisher/distributor that is not [normally] supplied to test users". If these contain valuable information, the overall evaluation should recommend that the publisher publishes these reports and/or make them available to test purchasers.

4. Information that is commercial in confidence:
   In some instances, publishers may have technically important material that they are unwilling to make public for commercial reasons. In practice there is very little protection available for intellectual property to test developers (copyright law being about the only recourse). Such information could include reports that cover the development of particular scoring algorithms, test or item generation procedures and report generation technology. Where the content of such reports might be important in making a judgment in a review, the association or organization responsible for the review could offer to undertake to enter into a non-disclosure agreement with the publisher. This agreement would be binding on the reviewers and editor. The reviewer could then evaluate the information and comment on the technical aspects and the overall evaluation to the effect that "the starred rating in this review refers to materials held by the publisher/ distributor that have been examined by the reviewers on a commercial in confidence basis. These are not supplied to end users."

Explanation of ratings

All sections are scored using the following rating system (see table on next page). Detailed descriptions giving anchor-points for each rating are provided.
Where a [0] or [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 only be used in exceptional circumstances by highly skilled experts or in research.
The instrument review needs to indicate which, given the nature of the instrument and its intended use, are the critical technical qualities. It is suggested that the convention to adopt is that ratings of these critical qualities are then shown in bold print.
In the following sections, overall ratings of the adequacy of information relating to validity, reliability and norms are shown, by default, in bold.

Any instrument with one or more [0] or [1] ratings regarding attributes that are regarded as critical to the safe use of that instrument, shall not be deemed to have met the minimum standard.

Rating	Explanation*
[n/a]	This attribute is not applicable to this instrument
0	Not possible to rate as no, or insufficient information is provided
1	Inadequate
2	Adequate
3	Good
4	Excellent

** A five point scale is defined by EFPA but each user can concatenate the points on the scale (for example combining points 3 and 4 into a single point). The only constraint is that there must be a distinction made between inadequate (or worse) on the one hand and adequate (or better) on the other. Descriptive terms or symbols such as stars or smiley faces may be used in place of numbers. Where the five point scale is replaced or customized, the user should provide a key that links the points and the nomenclature to the five point scale of EFPA.*

7 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 aspects or 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: e.g. the administrator's manual; technical handbooks; booklets of norms; manual supplements; updates from publishers/suppliers and so on.

Suppliers are asked to provide a complete set of such materials for each Reviewer. If you think there is something which users are supplied with which is not contained in the information sent to you for review, please contact your review editor.

7.1 Quality of the explanation of the rationale

If the instrument is a computer-adaptive test particular attention should be paid to the items 7.1.1 to 7.1. 6.

Items to be rated n/a or 0 to 4		**Rating**					
7.1.1	Theoretical foundations of the constructs	n/a	0	1	2	3	4
7.1.2	Test development (and/or translation or adaptation) procedure	n/a	0	1	2	3	4
7.1.3	Thoroughness of the item analyses and item analysis model	n/a	0	1	2	3	4
7.1.4	Presentation of content validity	n/a	0	1	2	3	4
7.1.5	Summary of relevant research	n/a	0	1	2	3	4
7.1.6	**Overall rating of the quality of the explanation of the rationale** This overall rating is obtained by using judgment based on the ratings given for items 7.1.1 – 7.1.5	n/a	0	1	2	3	4

7.2 Adequacy of documentation available to the user (user and technical manuals, norm supplements, etc.)

The focus here is on the quality of coverage provided in the documentation accessible to qualified users. Note that sub-section 7.2 is about the comprehensiveness and clarity of the documentation available to the user (user and technical manuals, norm supplements etc.) in terms of its coverage and explanation. In terms of the quality of the instrument as evidenced by the documentation, areas in this part are elaborated on under: 7.1, 7.3, 9, 10 and 11.

Items to be rated n/a or 0 to 4, 'benchmarks' are provided for an 'excellent' (4) rating.		Rating					
7.2.1	Rationale (see rating 7.1.6) Excellent: Logical and clearly presented description of what it is designed to measure and why it was constructed as it was.	n/a	0	1	2	3	4
7.2.2.1	Development Excellent: Full details of item sources, development of stimulus material according to accepted guidelines (e.g. Haladyna, Downing, & Rodriguez, 2002; Moreno, Martinez, & Muñiz, 2006), piloting, item analyses, comparison studies and changes made during development trials.	n/a	0	1	2	3	4
7.2.2.2	Development of the test through translation/adaptation Excellent: Information in the manual showing that the translation/adaptation process was done according to international guidelines (ITC, 2000) and included: • Input from native speakers of new language • Multiple review by both language and content (of test) experts • Back translation from new language into original language • Consideration of cultural and linguistic differences.	n/a	0	1	2	3	4
7.2.3	Standardisation Excellent: Clear and detailed information provided about sizes and sources of standardisation sample and standardisation procedure.	n/a	0	1	2	3	4
7.2.4	Norms Excellent: Clear and detailed information provided about sizes and sources of norms groups, representativeness, conditions of assessment etc.	n/a	0	1	2	3	4
7.2.5	Reliability Excellent: Excellent explanation of reliability and standard error of measurement (SEM), and a comprehensive range of internal consistency, temporal stability and/or inter-scorer and inter-judge reliability measures and the resulting SEM's provided with explanations of their relevance, and the generalisability of the assessment instrument.	n/a	0	1	2	3	4
7.2.6	Construct validity Excellent: Excellent explanation of construct validity with a wide range of studies clearly and fairly described.	n/a	0	1	2	3	4
7.2.7	Criterion validity Excellent: Excellent explanation of criterion validity with a wide range of studies clearly and fairly described.	n/a	0	1	2	3	4

7.2.8	Computer generated reports Excellent: Clear and detailed information provided about the format, scope, reliability and validity of computer generated reports.						
7.2.9	**Adequacy of documentation available to the user (user and technical manuals, norm supplements, etc.)** This rating is obtained by using judgment based on the ratings given for items 7.2.1 – 7.2.8	n/a	0	1	2	3	4

7.3 Quality of the procedural instructions provided for the user

Items to be rated n/a or 0 to 4, 'benchmarks' are provided for an 'excellent' (4) rating		**Rating**					
7.3.1	For test administration Excellent: Clear and detailed explanations and step-by-step procedural guides provided, with good detailed advice on dealing with candidates' questions and problem situations.	n/a	0	1	2	3	4
7.3.2	For test scoring Excellent: Clear and detailed information provided, with checks described to deal with possible errors in scoring. If scoring is done by the computer, is there evidence that the scoring is done correctly?	n/a	0	1	2	3	4
7.3.3	For norming Excellent: Clear and detailed information provided, with checks described to deal with possible errors in norming. If norming is done by the computer, is there evidence that score transformation is correct and the right norm group is applied?	n/a	0	1	2	3	4
7.3.4	For interpretation and reporting Excellent: Detailed advice on interpreting different scores, understanding normative measures and dealing with relationships between different scales, with illustrative examples and case studies; also advice on how to deal with the possible influence of inconsistency in answering, response styles, faking, etc.	n/a	0	1	2	3	4

7.3.5	For providing feedback and debriefing test takers and others Excellent: Detailed advice on how to present feedback to candidates including the use of computer generated reports (if available)	n/a	0	1	2	3	4
7.3.6	For providing good practice issues on fairness and bias Excellent: Detailed information reported about gender and ethnic bias studies, with relevant warnings about use and generalisation of validities	n/a	0	1	2	3	4
7.3.7	Restrictions on use Excellent: Clear descriptions of who should and who should not be assessed, with well-explained justifications for restrictions (e.g. types of disability, literacy levels required etc.)	n/a	0	1	2	3	4
7.3.8	Software and technical support Excellent: In the case of Computer Based Testing (CBT) or Web Based Testing (WBT): the information with respect to browser requirements, the installation of any required computer software and the operation of the software is complete (also covering possible errors and different systems), and availability of technical support is clearly described.	n/a	0	1	2	3	4
7.3.9	References and supporting materials Excellent: Detailed references to the relevant supporting academic literature and cross-references to other related assessment instrument materials.	n/a	0	1	2	3	4
7.3.10	**Quality of the procedural instructions provided for the user** This overall rating is obtained by using judgment based on the ratings given for items 7.3.1 – 7.3.9	n/a	0	1	2	3	4

7.4	**Overall adequacy** This overall rating for section 7 is obtained by using judgment based on the overall ratings given for the subsections 7.1, 7.2, and 7.3.	n/a	0	1	2	3	4

Reviewers' comments on the documentation: (comment on rationale, presentation and information provided)

8 Quality of the test materials

8.1 Quality of the test materials of paper-and-pencil tests

(this sub-section can be skipped if not applicable)

Items to be rated n/a or 0 to 4		Rating					
8.1.1	General quality of test materials (test booklets, answer sheets, test objects, etc.)	n/a	0	1	2	3	4
8.1.2	Ease with which the test taker can understand the task	n/a	0	1	2	3	4
8.1.3	Clarity and comprehensiveness of the instruction (including sample items and practice trials) for the test taker	n/a	0	1	2	3	4
8.1.4	Ease with which responses or answers can be made by the test taker	n/a	0	1	2	3	4
8.1.5	Quality of the formulation of the items and clarity of graphical content in the case of non-verbal items.	n/a	0	1	2	3	4
8.1.6	**Quality of the materials of paper-and-pencil tests** This overall rating is obtained by using judgment based on the ratings given for items 8.1.1 – 8.1.5	n/a	0	1	2	3	4

8.2 Quality of the test materials of CBT and WBT

(this sub-section can be skipped if not applicable)

Items to be rated n/a or 0 to 4		Rating					
8.2.1	Quality of the design of the software (e.g. robustness in relation to operation when incorrect keys are pressed, internet connections fail etc.)	n/a	0	1	2	3	4
8.2.2	Ease with which the test taker can understand the task	n/a	0	1	2	3	4
8.2.3	Clarity and comprehensiveness of the instructions (including sample items and practice trials) for the test taker, the operation of the software and how to respond if the test is administered by computer	n/a	0	1	2	3	4
8.2.4	Ease with which responses or answers can be made by the test taker	n/a	0	1	2	3	4
8.2.5	Quality of the design of the user interface	n/a	0	1	2	3	4
8.2.6	Security of the test against unauthorized access to items or to answers	n/a	0	1	2	3	4
8.2.7	Quality of the formulation of the items and clarity of graphical content in the case of non-verbal items.	n/a	0	1	2	3	4

8.2.8	**Quality of the materials of CBT and WBT** This overall rating is obtained by using judgment based on the ratings given for items 8.2.1 – 8.2.7	n/a	0	1	2	3	4

Reviewers' comments on quality of the materials

9 Norms

General guidance on assigning ratings for this section

It is difficult to set clear criteria for rating the technical qualities of an instrument. These notes provide some guidance on the sorts of values to associate with inadequate, adequate, good and excellent ratings. However these are intended to act as guides only. The nature of the instrument, its area of application, the quality of the data on which norms are based, and the types of decisions that it will be used for should all affect the way in which ratings are awarded.

To give meaning to a raw test 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 (see sub-section 9.1). 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)(see sub-section 9.2). With the latter two possibilities raw scores will be categorized in two (for example 'pass' of 'fail') or more different score ranges, e.g. 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.

9.1 Norm-referenced interpretation

(This sub-section can be skipped if not applicable)

Notes on international norms

Careful consideration needs to be given to the suitability of international (same language) norms. Where these have been carefully established from samples drawn from a group of countries, they should be rated on the same basis as nationally based (single language) norm groups. Where a non-local norm is provided strong evidence of equivalence for both test versions and samples to justify its use should be supplied. Generally such evidence would require studies demonstrating scalar equivalence between the source and target language versions. *Where this has not been reported then it should be commented upon in the Reviewers' comments at the end of section 9.*

An international norm may be the most appropriate for international usage (i.e. comparing people who have taken the test in different languages) but the issues listed below should be considered in determining its appropriateness. In general, use of an international norm requires the demonstration of at least measurement equivalence between the source and target language versions of the test.

The nature of the sample

- The balance of sources of the sample (e.g. a sample that is 95% German with a 2% Italian and 3% British is not a real international sample). A sample could be weighted to better reflect its different constituents.
- The equivalence of the background (employment, education, circumstances of testing etc.) of the different parts of the sample. Norm samples which do not allow this to be evaluated are insufficient.

The type of measure:

- Where there are measures which have little or no verbal content then there will be less impact on translation. This will apply to performance tests and to some extent to abstract and diagrammatic reasoning tests where should be less impact on the scores.

The equivalence of the test version used with the different language samples.

- There should be evidence that all the language versions are well translated/adapted
- Is there any evidence that any of the groups have completed the test in a non-primary language?

Similarities of scores in different samples:

- Evidence should be provided about the relative score patterns of the sample sections from different countries. Where there are large differences these should be accounted for and the implications in use discussed. E.g. if a Spanish sample scores higher on a scale than a Dutch sample is there an explanation of what it means to compare members of either group, or a third group against the average? Is there an interpretation of the difference?

Absence of these sources of evidence need to be commented upon in the Reviewers Comments at the end of the section

Guidance given about generalising the norms beyond those groups included in the international norms should be included in the manual for the instrument

- e.g. if a norm is made up of 20% German, 20% French,20% Italian, 20% British and 20% Dutch, it might be appropriate to use it as a comparison group for Swiss or Belgian candidates but it may not be appropriate to use it as a comparison for a group of Chinese applicants.

9.1	**Norm-referenced interpretation** Where an instrument is designed for use without recourse to norms or reference groups (e.g., ipsative tests designed for intra-individual comparisons only), the 'not applicable' category should be used rather than 'no information given'. However, the reviewer should evaluate whether the reasoning to provide no norms is justified, otherwise the category 'no information given' must be used.	
9.1.1	Appropriateness for local use, whether local or international norms Note that for adapted tests only local (nationally based) or really international norms are eligible for the ratings 2, 3 or 4 even if construct equivalence across cultures is found. Where measurement invariance issues arise separate norms should be provided for (sub)groups and any issues encountered should be explained.	
	Not applicable	n/a
	No information given	0
	Not locally relevant (e.g. inappropriate foreign samples)	1
	Local sample(s) that do(es) not fit well with the relevant application domain but could be used with caution	2
	Local country samples or relevant international samples with good relevance for intended application	3
	Local country samples or relevant international samples drawn from well-defined populations from the relevant application domain	4
9.1.2	Appropriateness for intended applications	
	Not applicable	n/a
	No information given	0

	Norm or norms not adequate for intended applications			1
	Adequate general population norms and/or range of norm tables, or adequate norms for some but not all intended applications			2
	Good range of norm tables			3
	Excellent range of sample relevant, age-related and sex-related norms with information about other differences within groups (e.g. ethnic group mix)			4
9.1.3	Sample sizes (classical norming) For most purposes, samples of less than 200 test takers will be too small, as the resolution provided in the tails of the distribution will be very small. The SE_{mean} for a *z*-score with $N = 200$ is 0.071 of the SD - or just better than one *T*-score point. Although this degree of inaccuracy may have only minor consequences in the centre of the distribution the impact at the tails of the distribution can be quite big (and this may be the score ranges that are most relevant for decisions to be taken on basis of the test scores). If there are international norms then in general, because of their heterogeneity, these need to be larger than the typical requirements of local samples. Different guideline figures are given for low and high stakes use. Generally high-stakes use is where a non-trivial decision is based at least in part on the test score(s).			
		Low-stakes use	High-stakes decisions	
	Not applicable			n/a
	No information given			0
	Inadequate sample size	e.g. < 200	e.g. 200-299	1
	Adequate sample size	e.g. 200-299	e.g. 300-399	2
	Good sample size	e.g. 300-999	e.g. 400-999	3
	Excellent sample size	e.g. ≥ 1000	e.g. ≥ 1000	4
9.1.4	Sample sizes continuous norming Continuous norming procedures have become more and more popular. They are used particularly for tests that are intended for use in schools (e.g. group 1 to 8 in primary education) or for a specific age range (e.g. an intelligence test for 6-16 year olds). Continuous norming is more efficient as fewer respondents are required to get the same amount of accuracy of the norms. Bechger, Hemker, and Maris (2009) have computed some values for the sizes of continuous norm groups that would give equal accuracy compared to classical norming. When eight subgroups are used $N = 70$ (8x70) gives equal accuracy compared to *N*s of 200 (8x200) with the classical approach; $N = 100$ (x8) compares to 300 (x8) and $N = 150$ (x8) to 400 (x8). In these cases the accuracy on the basis of the continuous norming approach is even better in the middle groups, but somewhat worse in the outer groups. Apart from the greater efficiency, another advantage is that, based on the regression line, values for intermediate norm groups can be computed. However, the approach is based on rather strict statistical assumptions. The test author has to show that these assumptions have been met, or that deviations from these assumptions do not have serious consequences for the accuracy of the norms. Note that when the number of groups is higher, the number of respondents in each group may be lower and vice versa. For high-stakes decisions, such as school admission, the required number shifts by one step upwards.			

	Not applicable	n/a
	No information given	0
	Inadequate sample size (e.g. fewer than 8 subgroups with a maximum of 69 respondents each)	1
	Adequate sample size (e.g. 8 subgroups with 70 - 99 respondents each)	2
	Good sample size (e.g. 8 subgroups with 100 - 149 respondents each)	3
	Excellent sample size (e.g. 8 subgroups with at least 150 respondents each)	4
9.1.5	Procedures used in sample selection (*select one*) A norm group must be representative of the referred group. A sample can be considered representative of the intended population if the composition of the sample with respect to a number of variables (e.g., age, gender, education) is similar to that of the population, *and* when the sample is gathered with a probability sampling model. In such a model the chance of being included in the sample is equal for each element in the population. In both probability and non-probability sampling different methods can be used. In probability sampling, when an individual person is the unit of selection, three methods can be differentiated: purely random, systematic (e.g. each tenth member of the population) and stratified (for some important variables, e.g. gender, numbers to be selected are fixed to guarantee representativeness on these variables). However (e.g. for the sake of efficiency), groups of persons can also be sampled (e.g. school classes), or a combination of group and individual sampling can be used. In non-probability sampling four methods are differentiated: pure convenience sampling (simply add every tested person to the norm group, as is done in most samples for personnel selection; post-hoc data may be classified into meaningful sub-groups based on biographical and situational information), quota sampling (as in convenience sampling, but it is specified before how many respondents in each subgroup are needed, as is done in survey research), snow ball sampling (ask you friends to participate, and ask them to ask their friends, etc.) and purposive sampling (e.g., select extreme groups to participate).	
	No information is supplied	[ ]
	Probability sample – random	[ ]
	Probability sample – systematic	[ ]
	Probability sample – stratified	[ ]
	Probability sample – cluster	[ ]
	Probability sample – multiphases (e.g. first cluster then random within clusters)	[ ]
	Non-probability sample – convenience	[ ]
	Non-probability sample – quota	[ ]
	Non-probability sample – ‘snow ball’	[ ]
	Non-probability sample – purposive	[ ]
	Other, describe:	[ ]

9.1.6	Representativeness of the norm sample(s)	
	Not applicable	n/a
	No information given	0
	Inadequate representativeness for the intended application domain or the representativeness cannot be adequately established with the information provided	1
	Adequate	2
	Good	3
	Excellent: Data are gathered by means of a random sampling model; 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; good representativeness with regard to these variables is established	4
9.1.7	Quality of information provided about minority/protected group differences, effects of age, gender etc.	
	Not applicable	n/a
	No information given	0
	Inadequate information	1
	Adequate general information, with minimal analysis	2
	Good descriptions and analyses of groups and differences	3
	Excellent range of analyses and discussion of relevant issues relating to use and interpretation	4
9.1.8	How old are the normative studies?	
	Not applicable	n/a
	No information given	0
	Inadequate, 20 years or older	1
	Adequate, norms between 15 and 19 years old	2
	Good, norms between 10 and 14 years old	3
	Excellent, norms less than 10 years old	4
9.1.9	Practice effects (only relevant for performance tests)	
	Not applicable	n/a
	No information given though practice effects can be expected	[ ]

	General information given	[ ]
	Norms for second test application after typical test-retest-interval	[ ]

9.2 Criterion-referenced interpretation

(This sub-section can be skipped if not applicable)

To determine the critical score(s) one can differentiate between procedures that make use of the judgment of experts (these methods are also referred to as domain-referenced norming, see sub-category 9.2.1) and procedures that make use of actual data with respect to the relation between the test score and an external criterion (referred to as criterion-referenced in the restricted sense, see sub-category 9.2.2).

9.2.1	**Domain-referenced norming**	
9.2.1.1	If the judgment of experts is used to determine the critical score, are the judges appropriately selected and trained? Judges should have knowledge of the content domain of the test and they should be appropriately trained in judging (the work of) test takers and in the use of the standard setting procedure applied. The procedure of the selection of judges and the training offered must be described.	
	Not applicable	n/a
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
9.2.1.2	If the judgment of experts is used to determine the critical score, is the number of judges used adequate? The required number of judges depends on the tasks and the contexts. The numbers suggested should be considered as an absolute minimum.	
	Not applicable	n/a
	No information given	0
	Inadequate (less than two judges)	1
	Adequate (two judges)	2
	Good (three judges)	3
	Excellent (four judges or more)	4

9.2.1.3	If the judgment of experts is used to determine the critical score, which standard setting procedure is reported? (*select one*)	
	Nedelsky	[ ]
	Angoff	[ ]
	Ebel	[ ]
	Zieky and Livingston (limit group)	[ ]
	Berk (contrast groups)	[ ]
	Beuk	[ ]
	Hofstee	[ ]
	Other, describe:	[ ]
9.2.1.4	If the judgment of experts is used to determine the critical score, which method to compute inter-rater agreement is reported? (*select one*)	
	Coefficient p_0	[ ]
	Coefficient Kappa	[ ]
	Coefficient Livingston	[ ]
	Coefficient Brennan and Kane	[ ]
	Intra Class Coefficient	[ ]
	Other, describe:	[ ]
9.2.1.5	If the judgment of experts is used to determine the critical score, what is the size of the inter-rater agreement coefficients (e.g. Kappa or ICC)? In the scientific literature there are no unequivocal standards for the interpretation of these kinds of coefficients, although generally values below .60 are considered insufficient. Below the classification of Shrout (1998) is followed. Using the classification needs some caution, because the prevalence or base rate may affect the value of Kappa.	
	Not applicable	n/a
	No information given	0
	Inadequate (e.g. $r < 0.60$)	1
	Adequate (e.g. $0.60 \leq r < 0.70$)	2
	Good (e.g. $0.70 \leq r < 0.80$)	3
	Excellent (e.g. $r \geq 0.80$)	4

9.2.1.6	How old are the normative studies?	
	Not applicable	n/a
	No information given	0
	Inadequate, 20 years or older	1
	Adequate, norms between 15 and 19 years old	2
	Good, norms between 10 and 14 years old	3
	Excellent, norms less than 10 years old	4
9.2.1.7	Practice effects (only relevant for performance tests)	
	No information given though practice effects can be expected	[ ]
	General information given	[ ]
	Norms for second test application after typical test-retest-interval	[ ]
9.2.2	**Criterion-referenced norming**	
9.2.2.1	If the critical score is based on empirical research, what are the results and the quality of this research? To answer this question no explicit guidelines can be given as to which level of relationship is acceptable, not only because what is considered 'high' or 'low' may differ for each criterion to be predicted, but also because prediction results will be influenced by other variables such as base rate or prevalence. Therefore, the reviewer has to rely on his/her expertise for his/her judgment. Also the composition of the sample used for this research (is it similar to the group for which the test is intended, more heterogeneous, or more homogeneous?) and the size of this group must be taken into account.	
	Not applicable	n/a
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
9.2.2.2	How old are the normative studies?	
	Not applicable	n/a
	No information given	0
	Inadequate, 20 years or older	1

	Adequate, norms between 15 and 19 years old	2
	Good, norms between 10 and 14 years old	3
	Excellent, norms less than 10 years old	4
9.2.2.3	Practice effects (only relevant for performance tests)	
	No information given though practice effects can be expected	[ ]
	General information given	[ ]
	Norms for second test application after typical test-retest-interval	[ ]
9.3	**Overall adequacy** This overall rating is obtained by using judgment based on the ratings given for items 9.1 – 9.2.2.3. The overall rating for *norm-referenced interpretation* can never be higher than the rating for the sample-size-item, but it can be lower dependent on the other information provided. From this other information especially information about the representativeness and the ageing of norms is relevant. If non-probability norm groups are used the quality of the norms can at most be qualified as 'adequate', but only when the description of the norm group shows that the distribution on relevant variables is similar to the target or referred group. The overall rating should reflect the characteristics of the largest and most meaningful norms rather than 'aver-age' across all published norms. The overall rating for *criterion-referenced interpretation* in case judges are used to determine the critical score can never be higher than the rating for the size of the inter-rater agreement, but it can be lower dependent on the other information provided. From this other information especially the correct application of the method concerned and the quality, the training and the number of judges are important. If the critical score is based on empirical research, the rating can never be higher than the rating for item 9.2.2.1, but it can be lower when the studies are too old.	
	Not applicable	n/a
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

Reviewers' comments on the norms: Brief report about the norms and their history, including information on provisions made by the publisher/author for updating norms on a regular basis. Comments pertaining to non-local norms should be made here.

10 Reliability

General guidance on assigning ratings for this section

Reliability refers to the degree to which scores are free from measurement error variance (i.e. a range of expected measurement error). For reliability, the guidelines are based on the need to have a small Standard Error for estimates of reliability. Guideline criteria for reliability are given in relation to two distinct contexts: the use of instruments to make decisions about groups of people (e.g. organizational diagnosis) and their use for making individual assessments. Reliability requirements are higher for the latter than the former. Other factors can also affect reliability requirements, such as the kind of decisions made and whether scales are interpreted on their own, or aggregated with other scales into a composite scale. In the latter case the reliability of the composite should be the focus for rating not the reliabilities of the components.

When an instrument has been translated and/or adapted from a non-local context, one could apply reliability evidence of the original version to support the quality of the translated/adapted version. In this case evidence of equivalence of the measure in a new language to the original should be proposed. Without this it is not possible to generalise findings in one country/language version to another. For internal consistency reliability evidence based on local groups is preferable, however, as this evidence is more accurate and usually easy to get. For some guidelines with respect to establishing equivalence see the introduction of the section on Validity. An aide memoire of critical points for comment when an instrument has been translated and/or adapted from a non-local context is included in the Appendix.

It is difficult to set clear criteria for rating the technical qualities of an instrument. These notes provide some guidance on the values to be associated with inadequate, adequate, good and excellent ratings. However these are intended to act as guides only. The nature of the instrument, its area of application, the quality of the data on which reliability estimates are based, and the types of decisions that it will be used for should all affect the way in which ratings are awarded. Under some conditions a reliability of 0.70 is fine; under others it would be inadequate. For these reasons, summary ratings should be based on your judgment and expertise as a reviewer and not simply derived by averaging sets of ratings.

In order to provide some idea of the range and distribution of values associated with the various scales that make up an instrument, enter the *number of scales* in each section. For example, if an instrument being used for *group-level decisions* had 15 scales of which five had retest reliabilities lower than 0.6, six between 0.60 and 0.70 and the other four in the 0.70 to 0.80 range, the median stability could be judged as 'adequate' (being the category in which the median of the 15 values falls). If more than one study is concerned, first the median value per scale should be computed, taking the sample sizes into account; in some cases results from a meta-analysis may be available, these can be judged in the same way. This would be entered as:

Stability	Number of *scales* *(if applicable)*	M*
No information given	[-]	0
Inadequate (e.g. $r < 0.60$)	[5]	1
Adequate (e.g. $0.60 \leq r < 0.70$)	[6]	2
Good (e.g. $0.70 \leq r < 0.80$)	[4]	3
Excellent (e.g. $r \geq 0.80$)	[0]	4

* M = median stability

For each of the possible ratings example values are given *for guidance only* - especially the distinctions between 'Adequate', 'Good' and 'Excellent'. For *high stakes decisions*, such as personnel selection, these

example values will be .10 higher. However, it needs to be noted that decisions are often based on aggregate scale scores. Aggregates may have much higher reliabilities than their component primary scales. For example, primary scales in a multi-scale instrument may have reliabilities around 0.70 while Big Five secondary aggregate scales based on these can have reliabilities in the 0.90s. Good test manuals will report the reliabilities of secondary as well as primary scales.

It is realised that it may be impossible to calculate actual median figures in many cases. What is required is your best estimate, given the information provided in the documentation. There is space to add comments. You can note here any concerns you have about the accuracy of your estimates. For example, in some cases, a very high level of internal consistency might be commented on as indicating a 'bloated specific'.

10	**Reliability**	
10.1	Data provided about reliability *(select two if applicable)*	
	No information given	[ ]
	Only one reliability coefficient given (for each scale or subscale)	[ ]
	Only one estimate of standard error of measurement given (for each scale or subscale)	[ ]
	Reliability coefficients for a number of different groups (for each scale or subscale)	[ ]
	Standard error of measurement given for a number of different groups (for each scale or subscale)	[ ]
10.2	**Internal consistency** The use of internal consistency coefficients is not sensible for assessing the reliability of speed tests, heterogeneous scales (also mentioned empirical or criterion-keyed scales; Cronbach, 1970), effect indicators (Nunnally & Bernstein, 1994) and emergent traits (Schneider & Hough, 1995). In these cases all items concerning internal consistency should be marked '*not applicable'*. It is also biased as a method for estimating reliability of ipsative scales. Alternate form or retest measures are more appropriate for these scale types. Internal consistency coefficients give a better estimate of reliability than split-half coefficients corrected with the Spearman-Brown formula. Therefore, the use of split-halves is only justified if, for any reason, information about the answers on individual items is not available. Split-half coefficients can be reported in item 10.7 (Other methods).	
10.2.1	Sample size	
	Not applicable	n/a
	No information given	0
	One inadequate study (e.g. sample size less than 100)	1
	One adequate study (e.g. sample size of 100-200)	2
	One large (e.g. sample size more than 200) or more than one adequate sized study	3
	Good range of adequate to large studies	4

10.2.2	Kind of coefficients reported *(select as many as applicable)*		
	Not applicable		n/a
	Coefficient alpha or KR-20		[ ]
	Lambda-2		[ ]
	Greatest lower bound		[ ]
	Omega (factor analysis)		[ ]
	Theta (factor analysis)		[ ]
	Other, describe: ……..		[ ]
10.2.3	Size of coefficients	Number of scales *(if applicable)*	M*
	Not applicable		n/a
	No information given	[ ]	0
	Inadequate (e.g. $r < 0.70$)	[ ]	1
	Adequate (e.g. $0.70 \leq r < 0.80$)	[ ]	2
	Good (e.g. $0.80 \leq r < 0.90$)	[ ]	3
	Excellent (e.g. $r \geq 0.90$)	[ ]	4
10.2.4	Reliability coefficients are reported with samples which …. (*select one*)		
	…. do not match the intended test takers, leading to more favourable coefficients (e.g. inflation by artificial heterogeneity)		[ ]
	…. do not match the intended test takers, but the effect on the size of the coefficients is unclear		[ ]
	…. do not match the intended test takers, leading to less favourable coefficients (e.g. reduction by restriction of range)		[ ]
	…. match the intended test takers		[ ]
	Not applicable		n/a
10.3	**Test retest reliability – temporal stability** Test retest refers to relatively short time intervals, whereas temporal stability refers to longer intervals in which more change is acceptable. Particularly for tests to be used for predictions over longer periods both aspects are relevant. To assess the temporal stability more than one retest may be required. The use of a test retest design is not sensible for assessing the reliability of state measures (actually a high test retest coefficient would invalidate the state character of a test). In this case all items concerning test retest reliability should be marked '*not applicable'*.		

10.3.1	Sample size		
	Not applicable		n/a
	No information given		0
	One inadequate study (e.g. sample size less than 100)		1
	One adequate study (e.g. sample size of 100-200)		2
	One large (e.g. sample size more than 200) or more than one adequate sized study		3
	Good range of adequate to large studies		4
10.3.2	Size of coefficients	Number of scales *(if applicable)*	M*
	Not applicable		n/a
	No information given	[ ]	0
	Inadequate (e.g. $r < 0.60$)	[ ]	1
	Adequate (e.g. $0.60 \leq r < 0.70$)	[ ]	2
	Good (e.g. $0.70 \leq r < 0.80$)	[ ]	3
	Excellent (e.g. $r \geq 0.80$)	[ ]	4
10.3.3	Data provided about the test-retest interval (*select or fill in test-retest interval*)		
	Not applicable		n/a
	No information given		[ ]
	The interval is:		……
10.3.4	Reliability coefficients are reported with samples which …. (*select one*)		
	…. do not match the intended test takers, leading to more favourable coefficients (e.g. inflation by artificial heterogeneity)		[ ]
	…. do not match the intended test takers, but effect on size of coefficients is unclear		[ ]
	…. do not match the intended test takers, leading to less favourable coefficients (e.g. reduction by restriction of range)		[ ]
	…. match the intended test takers		[ ]
	Not applicable		n/a

10.4	**Equivalence reliability (parallel or alternate forms)**		
10.4.1	Sample size		
	Not applicable		n/a
	No information given		0
	One inadequate study (e.g. sample size less than 100)		1
	One adequate study (e.g. sample size of 100-200)		2
	One large (e.g. sample size more than 200) or more than one adequate sized study		3
	Good range of adequate to large studies		4
10.4.2	Are the assumptions for parallelism* met for the different versions of the test for which equivalence reliability is investigated? *Note that tests can be considered to be parallel tests if in the same group the mean scores, variances and correlations with other tests are the same.		
	Not applicable		n/a
	No information given		0
	Inadequate		1
	Adequate		2
	Good		3
	Excellent		4
10.4.3	Size of coefficients	Number of scales *(if applicable)*	M*
	Not applicable		n/a
	No information given	[ ]	0
	Inadequate (e.g. $r < 0.70$)	[ ]	1
	Adequate (e.g. $0.70 \leq r < 0.80$)	[ ]	2
	Good (e.g. $0.80 \leq r < 0.90$)	[ ]	3
	Excellent (e.g. $r \geq 0.90$)	[ ]	4
10.4.4	Reliability coefficients are reported with samples which …. (*select one*)		
	…. do not match the intended test takers, leading to more favourable coefficients (e.g. inflation by artificial heterogeneity)		[ ]
	…. do not match the intended test takers, but effect on size of coefficients is unclear		[ ]

	…. do not match the intended test takers, leading to less favourable coefficients (e.g. reduction by restriction of range)	[ ]
	…. match the intended test takers	[ ]
	Not applicable	n/a
10.5	**IRT based method**	
10.5.1	Sample size It is difficult to give uniform guidelines for the adequacy of sample sizes in case IRT methods for the estimation of reliability are used, because the requirements are different in function of the item response format and the item response model used. Dependent on the item response model used minimum values for 'adequate' sample sizes are: 200 for 1-parameter studies, 400 for 2-parameter studies, and 700 for 3-parameter studies (based on Parshall, Davey, Spray, & Kalohn, 2001). These values apply to dichotomous models, but can be of some guidance for the reviewer when polytomous models are used for which the sample sizes may be smaller.	
	Not applicable	n/a
	No information given	0
	One inadequate study	1
	One adequate study	2
	One large or more than one adequate sized study	3
	Good range of adequate to large studies	4
10.5.2	Kind of coefficients reported (*select as many as applicable*) The first method gives the reliability of the estimated latent trait which in IRT replaces the estimated true score, i.e. test score (see Embretson & Reise, 2000). The second method is based on information about the individual items and gives an estimate of the reliability when the requirements typical for IRT are met (Mokken, 1971). The third method gives an estimate of the accuracy of the measurement related to the position on the latent trait.	
	Reliability of the estimated latent trait	[ ]
	Rho	[ ]
	Information function	[ ]
	Others, describe:	[ ]
	Not applicable	n/a

10.5.3	Size of coefficients (based on the final test length) Both guidelines for reliability coefficients (including rho) as for the information function are given. The guidelines for the information function are based on those for reliability coefficients since Information = $1/SE^2$, and given some often made assumptions, $r = 1 - SE^2$. Note that SE and information values are dependent on the value of the latent trait and that each test has a range within which the information value is optimal. The rating should not a priori be based on this optimal value, but on the information value of the score or range of scores that are of specific importance (e.g., critical scores). For these scores the information value may be optimal, but not necessarily so. If there are no such scores, the rating should be based on the mean information value (see also Reise & Havilund, 2005). Because there is not much experience with these rules-of-thumb, we advise raters to use these rules with care.	Number of scales *(if applicable)*	M*
	Not applicable		n/a
	No information given	[ ]	0
	Inadequate (e.g. $r < 0.70$; information < 3.33)	[ ]	1
	Adequate (e.g. $0.70 \leq r < 0.80$; $3.33 \leq$ information < 5.00)	[ ]	2
	Good (e.g. $0.80 \leq r < 0.90$; $5.00 \leq$ information < 10.00)	[ ]	3
	Excellent (e.g. $r \geq 0.90$; information ≥ 10.00)	[ ]	4
10.6	**Inter-rater reliability** If the scoring of a test involves no judgmental processes (e.g. simply summing the scores of multiple-choice items), this type of reliability is not required and all items concerning inter-rater reliability should be marked '*not applicable'*. Note that although inter-rater reliability may not apply to the test as a whole, it may apply to one or more subtests (e.g. some subtests of an intelligence test).		
10.6.1	Sample size		
	Not applicable		n/a
	No information given		0
	One inadequate study (e.g. sample size less than 100)		1
	One adequate study (e.g. sample size of 100-200)		2
	One large (e.g. sample size more than 200) or more than one adequate sized study		3
	Good range of adequate to large studies		4

10.6.2	Kind of coefficients reported (*select as many as applicable*)		
	Not applicable		n/a
	Percentage agree		[ ]
	Coefficient Kappa		[ ]
	Intra Class Correlation		[ ]
	Coefficient Iota		[ ]
	Other, describe:		[ ]
10.6.3	Size of coefficients To some methods mentioned in 10.6.2 the guide numbers may not apply as no *r*'s are computed.	Number of scales *(if applicable)*	M*
	Not applicable		n/a
	No information given	[ ]	0
	Inadequate (e.g. $r < 0.60$)	[ ]	1
	Adequate (e.g. $0.60 \leq r < 0.70$)	[ ]	2
	Good (e.g. $0.70 \leq r < 0.80$)	[ ]	3
	Excellent (e.g. $r \geq 0.80$)	[ ]	4
10.7	**Other methods of reliability estimation**		
10.7.1	Sample size		
	Not applicable		n/a
	No information given		0
	One inadequate study (e.g. sample size less than 100)		1
	One adequate study (e.g. sample size of 100-200)		2
	One large (e.g. sample size more than 200) or more than one adequate sized study		3
	Good range of adequate to large studies		4
10.7.2	Describe method:		
10.7.3	Results	Number of scales *(if applicable)*	M*
	Not applicable		n/a
	No information given	[ ]	0

	Inadequate	[ ]	1
	Adequate	[ ]	2
	Good	[ ]	3
	Excellent	[ ]	4

10.8	**Overall Adequacy** This overall rating is obtained by using judgment based on the ratings given for items 10.1 – 10.7.3. *Do not simply average numbers to obtain an overall rating.* For some instruments, internal consistency may be inappropriate (broad traits or scale aggregates), in which case more emphasis on the retest data should be placed. In other cases (state measures), retest reliabilities would be inappropriate, so emphasis should be placed on internal consistencies. For your final judgment you should also take into account: – whether the test is used for individual assessment or to make decisions on groups of people – the nature of the decision (high-stakes vs. low-stakes) – whether one or more (types of) reliability studies are reported – whether also standard errors of measurement are provided – procedural issues, e.g. group size, number of reliability studies, heterogeneity of the group(s) on which the coefficient are computed, number of raters if inter-rater agreement is computed, length of the test-retest interval, etc. – comprehensiveness of the reporting on the reliability studies.	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

Reviewers' comments on Reliability: Underline the strong and weak aspects of the evidence of reliability available. Comments pertaining to equivalence/reliability generalisation should also be made here (if applicable).

11 Validity

General guidance on assigning ratings for this section

Validity is the extent to which a test serves its purpose: can one draw the conclusions from the test scores which one has in mind? In the literature many types of validity are differentiated, e.g. Drenth and Sijtsma (2006, p. 334 – 340) 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 ways 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 variation in the attribute causally produces variation in the measured outcomes. Although this is a different approach, also in the opinion of these authors a differentiation between types a validity is not relevant.

However, whichever approach to validity one prefers, for a standardised judgment it is necessary to structure the concept of validity a bit. For this reason, separate sub-sections on construct and criterion validity are differentiated. Depending on the purpose of the test one of these aspects of validity may be more relevant than the other. However, it is realized 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 realized also, 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 one quality judgment is made about *the* (construct or criterion) validity of a test. This judgment should be a reflection of the quality of the evidence supporting the claim that the test can be used for the 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 rating 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).

When an instrument has been translated and/or adapted from a non-local context, evidence of equivalence of the measure in a new language to the original should be proposed. Without this it is not possible to generalise findings in one country/language version to another. Examples of equivalent evidence:

- Invariance in construct structure – e.g. via factor structure or correlation with standard measures.
- Similar criterion related validity – e.g. similar profile of correlations of a multi-scale instrument with independent external criterion – such as ratings of job competencies.
- Items show similar patterns of scale loadings e.g. items correlate in same pattern with other scales; strongest/weakest loading items are similar in original and new languages.
- Bilingual candidates have similar profiles in two languages (c.f. alternate form reliability).

Validity generalisation needs stronger evidence when translating tests across linguistic families (e.g. from an Indo-European to a Semitic language). In such a situation equivalence is under greater threat because of the differences in language structure and cultural differences. However, validity generalisation might be inferred from evidence of validity invariance in previous translations when a test has been translated into multiple languages. For instance, if a Swedish test has already been translated into French, German and Italian and has been shown to have equivalence in these languages.
In considering the whole issue of equivalence, it may be useful to follow Van de Vijver and Poortinga's (2005) classification:

- Structural / functional equivalence

  - There is evidence that the source and target language versions measure the same psychological constructs across groups. This is generally demonstrated by showing that patterns of correlations between variables are the same across groups.
- Measurement unit equivalence
  - There is evidence that the measurement units are the same, but there are different origins across groups (i.e. individual differences found in group A can be compared with differences found in group B, but the absolute raw scores for A and B are not directly comparable without some form of re-scaling).
- Scalar / Full score equivalence
  - The same measurement unit and the same origin (i.e. raw scores have the same meanings and can be compared across groups).

The benchmarks and the notes in the sub-sections 11.1 and 11.2 provide some guidance on the values to be associated with inadequate, adequate, good and excellent ratings. However these are intended to act as guides only. The nature of the instrument, its area of application, the quality of the data on which validity estimates are based, and the types of decisions that it will be used for should all affect the way in which ratings are awarded. For validity, guidelines on sample sizes are based on power analysis of the sample sizes needed to find moderate sized validities if they exist.

11.1 Construct validity

The purpose of construct validation is to find an answer to the question whether the test actually measures the intended construct or, partly or mainly, something else. Common methods for the investigation of construct validity are exploratory or confirmatory factor analysis, item-test correlations, comparison of mean scores of groups for which score differences may be expected, testing for invariance of factor structure and item-bias (DIF) for different groups, correlations with other instruments which are intended to measure the same (convergent validity) or different constructs (discriminant validity), Multi-Trait-Multi-Method research (MTMM), IRT-methodology and (quasi-)experimental designs.

11.1	**Construct validity**	
11.1.1	Designs used *(select as many as are applicable)*	
	No information is supplied	[ ]
	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	[ ]
	MTMM correlations	[ ]

	IRT methodology	[ ]
	(Quasi-)Experimental Designs	[ ]
	Other, describe:	[ ]
11.1.2	Do the results of (exploratory or confirmatory) factor analysis support the structure of the test?	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
11.1.3	Do the items correlate sufficiently well with the (sub)test score? Note that very high correlations may mean that items are more or less synonymous and that the concept measured may be very narrow (a so-called 'bloated specific')	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
11.1.4	Is the factor structure invariant across groups and/or is the test free of item-bias (DIF)? This kind of research can be carried out on basis of models within classical test theory or the IRT framework. If item-bias is found, the effect on the total score should be estimated (small effects are acceptable).	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

11.1.5	Are differences in mean scores between relevant groups as expected? E.g. pupils in group 8 are expected to score higher than pupils in group 6 on a test for numerical proficiency; children with the diagnosis ADHD should score higher on a test for hyperactivity than children not diagnosed with ADHD; salespersons should score higher on a test for commercial knowledge than the average working population. Even though the results are in the expected direction, this kind of research usually is inconclusive with respect to the construct validity of the test. However, the value of this kind of research is that when the expected differences are not shown, this would raise strong doubts about the construct validity of the test.	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
11.1.6	Median and range of the correlations between the test and tests measuring similar constructs An essential element of the process of construct validation is correlating the test score(s) with scales from similar instruments, the so-called congruent or convergent validity. The guidelines on congruent validity coefficients need to be interpreted flexibly. Where two very similar instruments have been correlated (with data obtained concurrently) we would expect to find correlations of 0.60 or more for 'adequate'. Where the instruments are less similar, or administration sessions are separated by some time interval, lower values may be adequate. When evaluating congruent validity, care should be taken when interpreting very high correlations. When correlations are above 0.90, the likelihood is that the scales in question are measuring exactly the same construct. This is not a problem if the scales in question represent a new scale and an established marker. It would be a problem though, if the scale(s) in question was (were) meant to be adding useful variance to what other scales already measure. The guidelines given concern correlations that are not adjusted for common-method variance or attenuation. Therefore, also the reliabilities of both instruments should be taken into account when judging the congruent validity coefficients. E.g., when both instruments have a reliability of .75, the maximum correlation between the instruments is .56. If reliabilities are higher, higher correlations are to be expected.	
	No information given	0
	Inadequate ($r < 0.55$)	1
	Adequate ($0.55 \leq r < 0.65$)	2
	Good ($0.65 \leq r < 0.75$)	3
	Excellent ($r \geq 0.75$)	4
11.1.7	Do the correlations with other instruments show good discriminant validity with respect to constructs that the test is not supposed to measure?	
	No information given	0
	Inadequate	1

	Adequate	2
	Good	3
	Excellent	4
11.1.8	If a Multi-Trait-Multi-Method design is used, do the results support the construct validity of the test (does it really measure what it is supposed to measure and not something else)? Note that if an MTMM design is used, research as mentioned in 11.1.6 and 11.1.7 may not be required anymore.	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
11.1.9	Other, e.g. IRT-methodology, (quasi-)experimental designs (describe):	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
11.1.10	Sample sizes The guidelines below concern studies within the classical test theory framework. For the estimation of item-parameters within IRT methodology 'adequate' sample sizes are: more than 200 for 1-parameter studies, more than 400 for 2-parameter studies and more than 700 for 3-parameter studies (based on Parshall, Davey, Spray, & Kalohn, 2001).	
	No information given	0
	One inadequate study (e.g. sample size less than 100)	1
	One adequate study (e.g. sample size of 100-200)	2
	One large (e.g. sample size more than 200) or more than one adequate sized study	3
	Good range of adequate to large studies	4
11.1.11	Quality of instruments as criteria or markers	
	No information given	0

	Inadequate quality	1
	Adequate quality	2
	Good quality	3
	Excellent quality with wide range of relevant markers for convergent and divergent validation	4
11.1.12	How old are the validity studies? It is difficult to formulate a general rule for taking the age of the research into account. For tests that intend to measure constructs in an area on which important theoretical developments have taken place, 15 year old research may be almost useless, whereas for other tests 20 year old (or even older) research still may be relevant.	
	Number of years	.....
11.1.13	**Construct validity - Overall adequacy** This overall rating is obtained by using judgment based on the ratings given for items 11.1.1 – 11.1.12. *Do not simply average numbers to obtain an overall rating.* In addition to the outcomes of the construct validity research, for your final judgment you should also take into account whether analysis techniques are used correctly (e.g. is the significance level corrected for correlating the instrument to other instruments without clear hypotheses, so-called 'fishing'), whether the research samples are similar to the group(s) for which the test is intended (e.g., more heterogeneity will inflate correlations, samples of students may give results that cannot be generalized), the size of the research sample(s), the quality of other instruments that are used (e.g. in convergent and discriminant validity research), and the age of the studies.	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

11.2 Criterion-related validity

Criterion-related evidence of validity (concurrent and predictive validity) refers to studies where real-world criterion measures (i.e. not other instrument scores) have been correlated with scales. Predictive studies generally refer to situations where assessment was carried out at a 'qualitatively' different point in time to the criterion measurement - e.g. for a work-related selection measure intended to predict job success, the instrument would have been carried out at the time of selection - rather than just being a matter of how long the time interval was between instrument and criterion measurement. Studies can also be 'post-dictive', for example, where scores on a potential selection test are correlated with job incumbents' earlier line manager ratings of performance. 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'.

11.2	**Criterion-related validity**	
11.2.1	Type of criterion study or studies (*select as many as are applicable)*	
	Predictive	[ ]
	Concurrent	[ ]
	Post-dictive	[ ]
11.2.2	Sample sizes	
	No information given	0
	One inadequate study (e.g. sample size less than 100)	1
	One adequate study (e.g. sample size of 100-200)	2
	One large (e.g. sample size more than 200) or more than one adequate sized study	3
	Good range of adequate to large studies	4
11.2.3	Quality of criterion measures	
	No information given	0
	Inadequate quality	1
	Adequate quality	2
	Good quality	3
	Excellent quality with respect to reliability and representation of the criterion construct	4
11.2.4	Strength of the relation between the test and criteria It is difficult to set clear criteria for rating the size of the criterion validity coefficients of an instrument. A criterion-related validity of 0.20 can have considerable utility in some situations, while one of 0.40 might be of little value in others. A coefficient of .30 may be considered good in personnel selection, whereas in educational situations higher coefficients are common. For these reasons, ratings should be based on your judgment and expertise as a reviewer and not simply derived by averaging sets of correlation coefficients. The guidelines given are based on Hemphill (2003; see also Meyer et al., 2001) and concern correlations that are not corrected for attenuation in either the predictor or the criterion. However, coefficients may be corrected for restriction of range.	

	The ranges given below concern validity coefficients, because *correlations* between tests and criteria are the most used way to represent criterion validity. 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. Swets (1988) presents an overview of values of ROC-curves in different areas. For certain types of medical diagnosis the values are between .81 and .97, for lie detection between .70 and .95, and for educational achievement (pass/fail) between .71 and .94. These values may be used as guidelines, but it is left to the expertise of the reviewer to decide to what extent the test can make a useful contribution to the decision concerned. Also when still other indices are reported, such as the positive and negative predictive value of a test, the likelihood ratio, etc.	
	No information given	0
	Inadequate ($r < 0.20$)	1
	Adequate ($0.20 \leq r < 0.35$)	2
	Good ($0.35 \leq r < 0.50$)	3
	Excellent ($r \geq 0.50$)	4
11.2.5	How old are the validity studies? It is difficult to formulate a general rule for taking the age of the research into account. For tests that intend to predict behaviour in rapidly changing environments, 15 year old research may be almost useless, whereas for other tests 20 year old (or even older) research may still be relevant.	
	Number of years	.....
11.2.6	**Criterion-related validity – Overall adequacy** This overall rating is obtained by using judgment based on the ratings given for items 11.2.1 – 11.2.5. *Do not simply average numbers to obtain an overall rating.* Apart from the outcomes of the criterion validity research, for your final judgment you should also take into account whether the right procedures and analysis techniques are used (e.g. is there criterion contamination, correction for attenuation, cross-validation), whether the research samples are similar to the group(s) for which the test is intended (e.g. correction for restriction of range), the size of the research sample(s), the quality of the criterion instruments that are used (e.g. is there criterion deficiency), and the age of the studies.	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

11.3 Overall validity

When judging overall validity, it is important to bear in mind the importance placed on construct validity as the best indicator of whether a test measures what it claims to measure. In some cases, the main evidence of this could be in the form of criterion-related studies. Such a test might have an 'adequate' or better rating for criterion-related validity and a less than adequate one for construct validity. In general the rating for Overall Validity will be equal to either the Construct Validity or the Criterion-related Validity, whichever is the greater. However, depending on the purpose of the test, one of these types of evidence may be considered more relevant than the other. The rating for Overall Validity should not be regarded as an average or as the lowest common denominator.

11.3	**Validity – Overall adequacy** This overall rating is obtained by using judgment based on the ratings given for items 11.1.1 – 11.2.6. *Do not simply average numbers to obtain an overall rating.*	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

Reviewers' comments on validity (all the evidence of validity included). Comments pertaining to equivalence/validity generalisation should also be made here (if applicable).

12 Quality of computer generated reports

Judging computer-based reports is made difficult by the fact that many suppliers will, understandably, wish to protect their intellectual property in the algorithms and scoring rules. In practice, sufficient information should be available for review purposes from the technical manual describing the development of the reporting process and its rationale, and through the running of a sample of test cases of score configurations. Ideally the documentation should also describe the procedures that were used to test the report generation for accuracy, consistency and relevance. For the purpose of reviewing at least three reports based on different score profiles including the actual scores should be provided, even if the algorithms for generating the reports are confidential.
For each of the following attributes, some questions are stated that should help you make a judgment, and a definition of an 'excellent' (4) rating is provided.

Items to be rated n/a or 0 to 4, 'benchmarks' are provided for an 'excellent' (4) rating.		
12.1	**Scope or coverage** 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. • *Does the report cover the range of attributes measured by the instrument?* • *Does it do so at a level of specificity justifiable in terms of the level of detail obtainable from the instrument scores?* • *Can the 'granularity' of the report (i.e. the number of distinct score bands on a scale that are used to map onto different text units used in the report) be justified in terms of the scales measurement errors?* • *Is the report designed for the same populations of people for whom the instrument was developed? (e.g. groups for whom the norm groups are relevant, or for whom there is relevant criterion data etc.).*	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent: Excellent fit between the scope of the instrument and the scope of the report, with the level of specificity in the report being matched to the level of detail measured by the scales. Good use made of all the scores reported from the instrument.	4
12.2	**Reliability** • *How consistent are the reports in their interpretation of similar sets of score data?* • *If report content is varied (e.g. by random selection from equivalent text units), is this done in a satisfactory manner?* • *Is the interpretation of scores and the differences between scores justifiable in terms of the scale measurement errors?*	

	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent: Excellent consistency in interpretation and appropriate warnings provided for statements, interpretation and recommendations regarding their underlying errors of measurement.	4
12.3	**Relevance or validity** The linkage between the instrument and the content of the report may be explained either within the report or be separately documented. Where reports are based on clinical judgment, the process by which the expert(s) produced the content and the rules relating scores to content should be documented. • *How strong is the relationship between the content of the report and the scores on the instrument? To what degree does the report go beyond or diverge from the information provided by the instrument scores?* • *Does the report content relate clearly to the characteristics measured by the instrument?* • *Does it provide reasonable inferences about criteria to which we might expect such characteristics to be related?* • *What empirical evidence is provided to show that these relationships actually exist?* It is relevant to consider both the construct validity of a report (i.e. the extent to which it provides an interpretation that is in line with the definition of the underlying constructs) and criterion-validity (i.e. where statements are made that can be linked back to empirical data).	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent: Relationship between the scales and the report content, with clear justifications provided.	4
12.4	**Fairness, or freedom from systematic bias** • *Is the content of the report and the language used likely to create impressions of inappropriateness for certain groups?* • *Does the report make clear any areas of possible bias in the results of the instrument?* • *Are alternate language forms available? If so, have adequate steps been taken to ensure their equivalence?*	
	No information given	0
	Inadequate	1
	Adequate	2

	Good	3
	Excellent: Clear warnings and explanations of possible bias, available in all relevant user languages.	4
12.5	**Acceptability** This will depend substantially on the complexity of the language used in the report, the complexity of the constructs being described and the purpose for which it is intended. • *Is the form and content of the report likely to be acceptable to the intended recipients?* • *Is the report written in a language that is appropriate for the likely levels of numeracy and literacy of the intended reader?*	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent: Very high acceptability, well-designed and well-suited to the intended audience.	4
12.6	**Length** This is also an aspect of Practicality and should be reflected in the rating given for this, but too long reports may also be an indication of over-interpretation of scores. Therefore the length of reports is rated separately also. Generally reports that on average take more than one page per scale (excluding title pages, copyright notices etc.) may be over long and over-interpreted.	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
12.7	**Overall adequacy of computer generated reports** This overall rating is obtained by using judgment based on the ratings given for items 12.1 –12.6. *Do not simply average numbers to obtain an overall rating.*	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

Reviewers' comments on computer generated reports

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.

Comments on the complexity of the algorithms can be included, e.g. whether multiple scales are considered simultaneously, how scale profiles are dealt with etc. Such complexity should, of course, be supported by a clear rationale in the manual.

13 Final evaluation

Evaluative report of the test

This section should contain a concise, clearly argued judgment about the test. It should describe its pros and cons, and give some general recommendations about how and when it might be used - together with warnings (where necessary) about when it should not be used.

A summary of any positive or negative points raised in connection with adapted and translated tests should be summarised here. A checklist of the important considerations for such instruments is added in the Appendix as a reminder of the notes in the relevant sections. Only comment on these if this is appropriate.

The evaluation should cover topics such as the appropriateness of the instrument for various assessment functions or areas of application; any special training needs or special skills required; whether training requirements are set at the right level; ease of use; the quality and quantity of information provided by the supplier and whether there is important information which is not supplied to users and where there are issues arising from the instrument being translated or adapted (see Appendix).

Include comments on any research that is known to be under way, and the supplier's plans for future developments and refinements etc.

Conclusions

Recommendations (*select one)* The relevant recommendation, from the list given, should be indicated. Normally this will require some comment, justification or qualification. A short statement should be added relating to the situations and ways in which the instrument might be used, and warnings about possible areas of misuse. **All the characteristics listed below should have ratings of either n/a, 2, 3, or 4 if an instrument is to be 'recommended' for general use (box 4 or 5)**. 9 Norms 10 Reliability−overall 11 Validity-overall 12 Computer generated reports If any of these ratings are 0 or 1 the instrument will normally be classified under Recommendation 1, 2, or 3 or it will be classified under 'Other' with a suitable explanation given.	1 Requires further development. Only suitable for use in research, not for use in practice	[ ]
	2 Only suitable for use by an expert user (exceeding EFPA User Qualification Level 2) under carefully controlled conditions or in very limited areas of application	[ ]
	3 Suitable for supervised use in the area(s) of application defined by the distributor by any user with general competence in test use and test administration (exceeding EFPA User Qualification Level 2)	[ ]
	4 Suitable for use in the area(s) of application defined by the distributor, by test users who meet the distributor's specific qualifications requirements (at least EFPA User Qualification Level 2)	[ ]
	5 Suitable for unsupervised self-assessment in the area(s) of application defined by the distributor	[ ]
	6 Other ……	[ ]

PART 3 BIBLIOGRAPHY

American Educational Research Association, American Psychological Association, and 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. (2002a). *EFPA Review Model for the description and evaluation of psychological instruments: Version 3.2. Evaluation Form.* Brussels: EFPA Standing Committee on Tests and Testing (September, 2002).

Bartram, D. (2002b). *EFPA Review Model for the description and evaluation of psychological instruments: Version 3.2. Notes for Reviewers*. Brussels: EFPA Standing Committee on Tests and Testing (September, 2002).

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.

Bartram, D., Lindley, P. A., & Foster, J. M. (1992). *Review of psychometric tests for assessment in vocational training*. BPS Books: Leicester.

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. (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.

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

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

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 in the theory and application of psychological tests (4th revised ed.)]. Houten, The Netherlands: Bohn Stafleu van Loghum.

Embretson, S. E. (Ed.) (2010*). Measuring psychological constructs. Advances in model-based approaches.* Washington, D. C.: American Psychological Association.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*.Mahwah, NJ: Erlbaum.

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., Braak, M., Frima, R., & van Vliet-Mulder, J. C. (2009-2012). *Documentatie van Tests en Testresearch in Nederland* [Documentation of Tests and Testresearch in The Netherlands]. Amsterdam: Boom test uitgevers.

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*, in press.

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.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309-334.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, *24*, 355–366.

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

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*, 78-80.

International Test Commission. (2005). *International Guidelines on Computer-Based and Internet Delivered Testing*. Bruxelles, Belgium: Author.

Kersting, M. (2008). DIN Screen, Version 2. Leitfaden zur Kontrolle und Optimierung der Qualität von Verfahren und deren Einsatz bei beruflichen Eignungsbeurteilungen [DIN Screen, Version 2. Guide line for monitoring and optimizing the quality of instruments and their application in proficiency assessment procedures.]. In M. Kersting. *Qualitätssicherung in der Diagnostik und Personalauswahl - der DIN Ansatz* (S. 141-210) [Guaranteeing quality in diagnostics and personnel selection (p. 141-210)]. Göttingen: Hogrefe.

Lindley, P. A. (2009). *Reviewing translated and adapted tests: Notes and checklist for reviewers:5 May 2009*. Leicester, UK: British Psychological Society. Retrieved from http://www.efpa. eu/professional-development/tests-and-testing.

Lindley, P.A. (2009, July). Using EFPA Criteria as a common standard to review tests and instruments in different countries. In D.Bartram (Chair), *National approaches to test quality assurance*. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

Lindley, P., Bartram, D., & Kennedy, N. (2004). *EFPA Review Model for the description and evaluation of psychological tests: test review form and notes for reviewers: Version 3.3.* Leicester, UK: British Psychological Society (November, 2004).

Lindley, P., Bartram, D., & Kennedy, N. (2005). *EFPA Review Model for the description and evaluation of psychological tests: test review form and notes for reviewers: Version 3.41.* Brussels: EFPA Standing Committee on Tests and Testing (August, 2005).

Lindley, P., 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.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128-165.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton.

Moosbrugger, H., Kelava, A., Hagemeister, C., Kersting, M., Lang, F., Reimann, G., et al. (2009, July). The German Test Review System (TBS-TK) and first experiences. In D. Bartram (Chair), *National approaches to test quality assurance*. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

Moreno, R., Martínez, R. J., & Muñiz, J. (2006). New guidelines for developing multiple-choice ítems. *Methodology, 2*, 65-72.

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

Nielsen, S. L. (2009, July). Test certification through DNV in Norway. In D. Bartram (Chair), *National approaches to test quality assurance*. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

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

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.

Reise, S. P., & Havilund, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Measurement, 84*, 228-238.

Tideman, E. (2007). Psychological tests and testing in Sweden. *Testing International*, *17*(June), 5–7.

Schneider, R. J., & Hough, L. M. (1995). Personality and industrial/organizational psychology. In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology, 10,* 75-129.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research, 7*, 301-317.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285-1293.

Testkuratorium. (2009). TBS-TK. Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 09. September 2009 [TBS-TK. Test review system of the board of testing of the Federation of German psychologists' associations]. *Report Psychologie, 34,* 470-478.

Van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Erlbaum.

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

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

Ziegler, M., MacCann, C., & Roberts, R. (Eds.) (2011). *New perspectives on faking in personality assessment*. Oxford, UK: Oxford University Press.

APPENDIX

An aide memoire of critical points for comment when an instrument has been translated and/or adapted from a non-local context

Development	
Evidence or discussion of	*Input from native speakers of new language*
	Multiple review by both language and content (of test) experts
	Back translation from new language into original language
Basic psychometric properties	*Item performance*
	Reliability
Norms	
	A local norm is provided
Non-local norm	*Strong evidence of equivalence for both test versions and samples*
International norms	*Larger than the typical requirements of local samples*
The nature of the sample	*Balance of sources of the sample*
	Equivalence of the background of the different parts of the sample
The type of measure	*Little or no verbal content*
The equivalence of the test version	*All the language versions are well translated/adapted*
	Some groups have completed the test in a non-primary language
Similarities of scores in different samples	*Where there are large differences these should be accounted for and the implications in use discussed*
Guidance about generalising the norms	
Equivalence/ Reliability/Validity	
Invariance in construct structure	*Via factor structure, equivalence of correlation matrices or similarity of patterns of correlation with standard measures*
Similar criterion related validity	*Strongest correlation with similar competencies*
Similar patterns of scale loadings	*Items correlate in same pattern with other scales*
	Strongest/weakest loading items are similar in original and new languages
Alternate form reliability	*Bilingual candidates have similar profiles in two languages*
Validity generalisation	
Validity generalisation needs strong evidence	*When translating tests across linguistic families (e.g. from an Indo-European to a Semitic language*
Validity generalisation can be inferred	*Where a test has been translated into multiple languages some validity generalisation can be inferred from evidence of validity invariance in previous translations: Swedish test has already been translated into French, German and Italian and has been shown to have equivalence in these languages*