efpa
European Federation of
Psychologists' Associations

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

TEST REVIEW FORM AND NOTES FOR REVIEWERS

VERSION 5.0

Version 5.0 is a major revision of Version 4.2.6 (2013) by a task force of the Board of Assessment of EFPA consisting of:

Ana Maria Hernandez Baeza (Spain), Helen Baron (EAWOP), Urszula Brzezinska (Poland), Iris Egberink (The Netherlands), Nigel Evans (UK), Ian Florance (ETPG), Steven Joris (Belgium), Dragos Illiescu (Romania), Mark Schittekatte (chair, Belgium)

Approved by EFPA Board of Directors MONTH / YEAR

Wording to be considered by EFPA Head Office in consultation with the organisations concerned

Users of this document and its contents are required by EFPA to acknowledge this source with the following text: "This version of the EFPA Test Review Criteria was based on the earlier approved version 4.2.6, which in turn was largely modelled on the form and content of the British Psychological Society's (BPS) test review criteria and criteria developed by the Dutch Committee on Tests and Testing (COTAN) of the Dutch Association of Psychologists (NIP) and the Spanish Questionnaire for the Evaluation of Psychometric Tests (developed by the Spanish Psychological Association). This new version has also drawn on the revision of COTAN's criteria. EFPA is grateful to the BPS, the Spanish Association and the NIP for permission to build on their criteria in developing the European model. All intellectual property rights in the original BPS and NIP criteria are acknowledged and remain with those bodies."

European Federation of Psychologists' Associations AISBL - EFPA
Rue Marché aux Herbes 105 / 39 – 1000 Bruxelles - Belgique
Grasmarkt 105 / 39 - 1000 Brussels – Belgium

Contents

Introduction

The EFPA Test Review Model provides a structure for descriptions and rigorous evaluations of psychological assessments, tests, scales, profiles, and questionnaires used in work, education, health, sports, forensic, counselling, coaching and other contexts.

Review information presented in this structure will support developers, authors, suppliers, publishers, and trainers to improve tests and testing practice. It will also inform policy makers in defining and supporting standards and, in particular, in creating a test review programme. In turn this will help users make the right assessment choices by providing authoritative, unbiased, consistent reviews of tests.

Following the Standards for Educational and Psychological Testing (2014) the term "test" is used for any "psychometrically derived measurement instrument that assesses the psychological constructs in which a structured sample of an examinee's behaviour in a specified domain is obtained and subsequently quantified, scored, interpreted, and synthesized using a standardized process for the purpose of evaluative conclusion or recommendation." The EFPA test review model can also be applied to instruments that measure groups of people (e.g. teams). It applies to all instruments that are covered by this definition, whether called a scale, questionnaire, projective technique, profile, structured interviewing system or structured life history.

This model is divided into three main parts. The first describes the test in detail. The second evaluates fundamental properties of the test, covering: test materials, norms, reliability, validity, its approach to *Equality, Diversity and Inclusion* (EDI) and computer-generated reports ending in a global evaluation. The final section provides references used in the review, a bibliography and glossary.

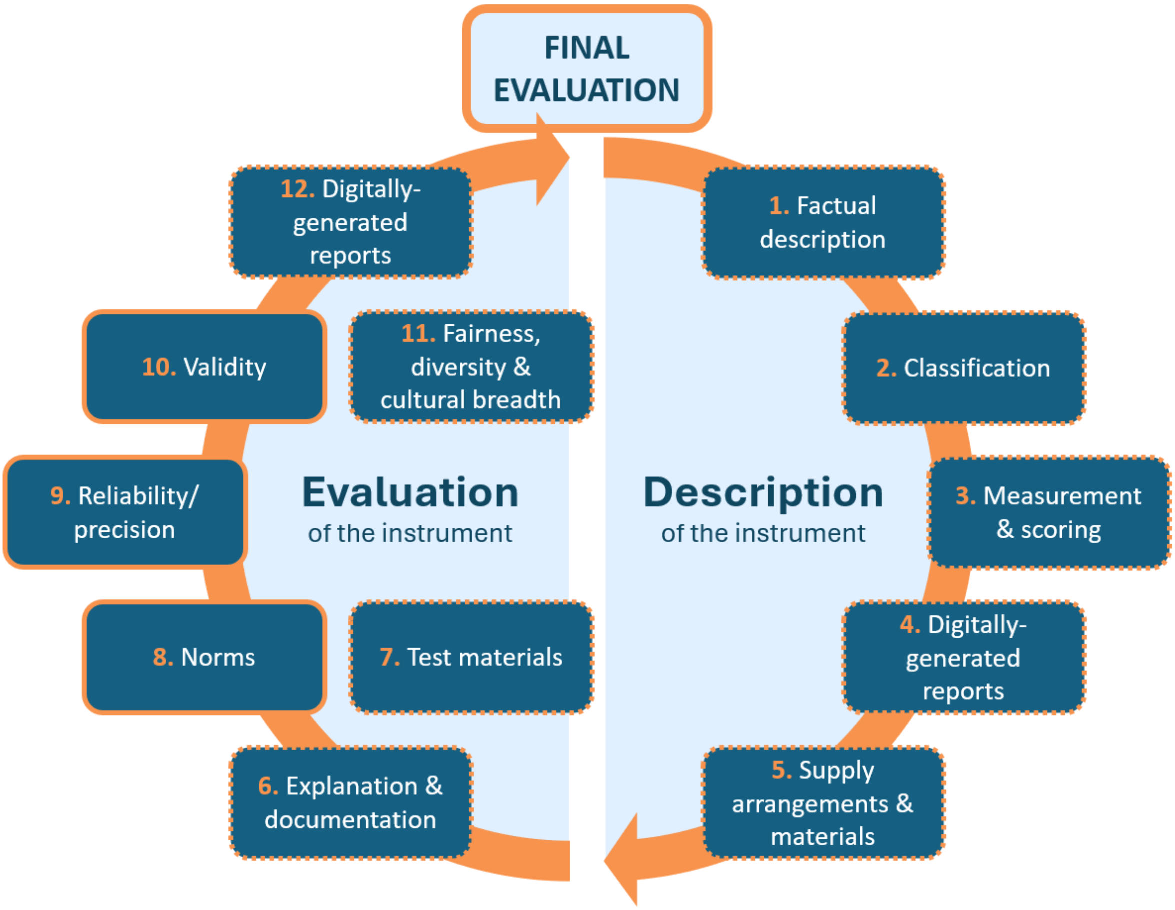
How the model should be used

Effective implementation is as important as the model itself. Earlier editions have been operationalized by a number of organizations, most specifically by national psychological associations on their web sites. Other organizations may want to use this model as the basis for review systems of tests used in their particular discipline/application in diverse media. It may also be used as an accreditation system if this is implemented by national bodies and regulators. The process of operationalising earlier versions has varied from country to country depending on local professional guidelines and laws.

Separate white papers describe how to apply to EFPA to use the model, and solutions used with earlier versions, these are available at EFPA's website. 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 model or in the review procedures when countries start to use the model. In addition, test developers and publishers are encouraged to use the model to evaluate the quality of their own tests.

This model is part of EFPA's information strategy: to evaluate instruments and the technical information supporting them. The star ratings given in this review do not designate EFPA's official approval or recommendation of a test; advertising by publishing companies and others must not state or imply that this is the case but should reference the model if it was used.

Comments on this model are welcomed in the hope that the experiences of users will be instrumental in improving and clarifying the processes, and should be addressed to EFPA using the contact information at the end of this document.


FINAL EVALUATION
1. Factual description
2. Classification
3. Measurement & scoring
4. Digitally-generated reports
5. Supply arrangements & materials
6. Explanation & documentation
7. Test materials
8. Norms
9. Reliability/ precision
10. Validity
11. Fairness, diversity & cultural breadth
12. Digitally-generated reports
Evaluation of the instrument
Description of the instrument

The use of bookmarks

This document consists of 3 parts. Each part is further divided into chapters. In an attempt to make this document as user-friendly as possible, bookmarks have been added so that users can easily navigate between the different parts.

- Clicking (first press Ctrl) on ‘Part 1’ of this document will automatically jump you to ‘Part 2’, and so on.
- Clicking (first press Ctrl) on any chapter title will automatically jump you to the next chapter title, and so on.

Note on this revision

This version of the model has been prepared by a working party of the EFPA Board of Assessment comprising Ana Maria Hernandez Baeza (Spain), Helen Baron (EAWOP), Urszula Brzezinska (Poland), Iris Egberink (The Netherlands), Nigel Evans (UK), Ian Florance (ETPG), Steven Joris (Belgium), Dragos Iliescu (Romania), Mark Schittekatte (chair, Belgium)

Since the last version of this model in 2013, test practice has become more diverse across Europe, reflecting developments in technology and psychology as well as the wider range of professions using tests. Different countries recognise different professional and/or competence requirements for the purchase and use of psychological tests in different domains. In certain assessment areas and different countries, delivery has largely migrated onto online environments; interpretation has become digitised; feedback often takes place at a distance.

This present version of the review model therefore aims to achieve several goals. It attempts to rebalance the emphasis between print and digital tests. How limited or widely test usage is allowed is obviously an issue for local organisations but this model seeks to address all test users to ensure testing practice is of a consistently high quality across Europe. More general changes in European society are reflected in the use of more inclusive language. New legal regulations relating to data, digital technology and medical devices may also impact testing.

Previous versions

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 were 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 last version of the model was prepared by a Task Force of the EFPA Board of Assessment, whose members were Arne Evers (Chair, the Netherlands), Carmen Hagemeister (Germany), Andreas Høstmælingen (Norway), Patricia Lindley (UK), José Muñiz (Spain), and Anders Sjöberg (Sweden). 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 were integrated. The texts of some major updated passages were based on the revised Dutch rating system for test quality (Evers, Lucassen, Meijer, & Sijtsma, 2010; Evers, Sijtsma, Lucassen, & Meijer, 2010).

Part 1. Description of the instrument

1. Factual description

General information	
Reviewer *(each country can decide either to publish the reviewers' names when the integrated review is published or to opt for anonymous reviewing)*	Click or tap here to enter text.
Date of current review	Click or tap here to enter text.
Date of previous review *(if applicable)*	Click or tap here to enter text.
Instrument name *(local version)*	Click or tap here to enter text.
Short name of the test *(if applicable)*	Click or tap here to enter text.
Original test name *(if the local version is an adaptation)*	Click or tap here to enter text.
Authors of the original test	Click or tap here to enter text.
Authors of the local adaptation	Click or tap here to enter text.
Local test distributor/publisher	Click or tap here to enter text.
Publisher of the original version of the test *(if different to current distributor/publisher)*	Click or tap here to enter text.
Date of publication of current revision/edition	Click or tap here to enter text.
Date of publication of adaptation for local use	Click or tap here to enter text.
Date of publication of original test	**Click or tap here to enter text.**

2. Classification

Unless otherwise indicated, select from those descriptions that the publisher provides. Where these are not clear, indicate this fact and judge from the information provided in the manual the most appropriate answers. Where the publishers' suggestions seem inappropriate, the evaluation part of the review should include comments.

2.1	**Content domains** Specify what the test measures using up to 3 keywords. ☐ Not explicitly stated ☐ Ability ☐ Attention ☐ Emotional Intelligence ☐ Group Function ☐ Interests ☐ IQ ☐ Learning ☐ Manual Dexterity ☐ Motivation ☐ Non-verbal ☐ Personality ☐ Potential ☐ Projective ☐ Scholastic attainment ☐ Sensorimotor ☐ Verbal ☐ Other *(describe)*: Click or tap here to enter text.
2.2	**Area of use** Select all that apply. ☐ Not explicitly stated ☐ Advice, guidance, and career choice ☐ Clinical ☐ Educational ☒ Forensic ☐ General health, life, and well-being ☐ Neurological ☐ Sports and Leisure ☐ Work and Organisational ☐ Other *(describe)*: Click or tap here to enter text.
2.3	**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. Click or tap here to enter text.
2.4	**Main intended users** Select all that apply. ☐ Not explicitly stated ☐ Clinical Psychologists ☐ Health Professionals ☐ HR Professionals ☐ Qualified Psychologists ☐ Specialist Teachers ☐ Speech and Language Therapists ☐ Other *(describe)*: Click or tap here to enter text.
2.5	**Number of scales and brief description of the variable(s) measured** Indicate the number of scales and provide a brief description of each if its meaning is not clear from its name. These should include 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 multi-trait personality test, or subtest, factor, and total scores on an intelligence test. Click or tap here to enter text.

2.6

Response mode

Select all that apply.

Describe any special pieces of equipment which are required if they are not included in the list of options opposite.

☐ Not explicitly stated ☐ Behavioural interaction ☐ Drawing ☐ Keyboard or mouse responses ☐ Manual (physical) operations ☐ Oral ☐ Paper and pencil ☐ Touch screen
☐ Specialist response device *(describe):* Click or tap here to enter text.
☐ Other *(describe):* Click or tap here to enter text.

2.7

Demands on the test taker

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 ability to complete a 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 capability is not necessary at all.
☐ "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 but this is not clear from information provided by the test publisher.

Select one classification in each case.

Attention
☐ irrelevant / not necessary
☐ necessary information given
☐ information missing

Command of test language (understanding and speaking)
☐ irrelevant / not necessary
☐ necessary information given
☐ information missing

Digital literacy /experience
☐ irrelevant / not necessary
☐ necessary information given
☐ information missing

Digital skills
☐ irrelevant / not necessary
☐ necessary information given
☐ information missing

Handedness
☐ irrelevant / not necessary
☐ necessary information given
☐ information missing

Hearing
☐ irrelevant / not necessary
☐ necessary information given
☐ information missing

Manual capabilities
☐ irrelevant / not necessary
☐ necessary information given
☐ information missing.

Reading
☐ irrelevant / not necessary
☐ necessary information given
☐ information missing

Vision
☐ irrelevant / not necessary
☐ necessary information given
☐ information missing

Writing
☐ irrelevant / not necessary
☐ necessary information given
☐ information missing

Other *(describe)*: Click or tap here to enter text.

2.8 **Special testing conditions**

Describe any specific testing conditions which may be required.

Click or tap here to enter text.

2.9 **Item response types**

Select all that apply.
For example: when the test uses multiple response types select all the types it uses.

☐ Not explicitly stated ☐ Graded scale ratings (e.g. Likert) ☐ Interactions/choices in computer generated environment ☐ Interactions/choices in real environment ☐ Multiple choice (ability testing, or right/wrong, yes/no) ☐ Multiple choice (mixed scale alternatives) ☐ Open ☐ Rankings ☐ Response latency ☐ Response times ☐ Task success in computer generated environment ☐ Task success in real environment.
☐ Other *(describe*): Click or tap here to enter text.

2.10 **Item stimulus type**

Select all that apply.

☐ Not explicitly stated ☐ Abstract images ☐ Closed questions ☐ Game environments ☐ Open questions ☐ Photographs ☐ Representative images ☐ Scenarios/case studies ☐ Sound clips ☐ Video clips
☐ Other *(describe*): Click or tap here to enter text.

2.11 **Total number of items and number of items per scale or subtest**

Where the test is not static, for example adaptive testing or gamified environments, indicate the minimum, maximum and typical number of items or measurement points.

Click or tap here to enter text.

2.12 **Intended mode of administration**

These should reflect the conditions under which the instrument was developed. Note that usage modes may vary across versions of a tool. Mark both if appropriate.

Suitable for:
☐ Individual administration ☐ Group administration

2.13 **Technological arrangements available/required to administer the test**

Mark A (available) / R (required) against each option.

☐ Not explicitly stated ☐ Paper and Pencil ☐ PC without connectivity ☐ PC with connectivity ☐ Phone without connectivity ☐ Phone with connectivity ☐ Proprietary apparatus ☐ Tablet without connectivity ☐ Tablet with connectivity

☐ Other (*describe*): Click or tap here to enter text.

2.14 **Time required to use the test**

In many cases, only general estimates of timing rather than precise figures will be possible. Where a function is automated, the required time is 0. Do not include the time needed to become familiar with the instrument itself. Assume the user is experienced and qualified.

- Preparation: 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.

  Click or tap here to enter text.
- Administration per session: 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. In much automated test administration, these elements will be performed by the system and the time required will be 0.

  Click or tap here to enter text.
- Scoring: the time taken to obtain raw-scores. This may be automated.

  Click or tap here to enter text.
- Analysis: the time taken to carry out further work on the raw scores to derive other measures and to produce a reasonably comprehensive interpretation. This may be automated.

  Click or tap here to enter text.
- Feedback: the time required to prepare and provide feedback to a test taker and other stakeholders. Where automatically generated reports are used, only list the time required to support understanding of these. This could be through the provision of a helpline in case of queries or could be a session to explain the report's findings and any time required to assimilate the report findings before such sessions.

  Click or tap here to enter text.

2.15 **Different forms**

Are there alternative versions (genuine or pseudo-parallel forms, short versions, etc?). If so, describe the applicability of each for different groups of people.

Some tests offer equivalent 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 versions). Also describe whether or not parts of the whole test can be used instead of the whole instrument.

Click or tap here to enter text.

2.16	**Static or dynamic determination** Where the test is not static, for example with adaptive testing or in gamified environments, how is the content the test taker receives determined? ☐ Not explicitly stated ☐ Static test form(s) ☐ Form adaptive to test takers’ responses from items generated on the fly ☐ Form adaptive to test takers’ responses generated from a fixed item pool ☐ Form adaptive to test takers scores from items generated on the fly ☐ Form adaptive to test takers’ scores generated from a fixed item pool ☐ Forms created from items generated on the fly ☐ Forms randomly generated from a fixed item pool

3. Measurement and scoring

3.1

Scoring procedure
Select all that apply.

☐ Not explicitly stated
☐ Digital scoring by Optical Mark Reader entry of responses from the paper response form
☐ Digital scoring with manual entry of responses from a paper response form
☐ Digital scoring from direct entry of responses by test taker
☐ Simple manual scoring key
☐ Complex manual scoring
☐ Bureau-service
☐ Other *(describe):* Click or tap here to enter text.

3.2

Scores
Brief description of the scoring system to obtain global and partial scores, (correction for guessing, qualitative interpretation aids, etc.).

Click or tap here to enter text.

3.3

Scales used
Select all that apply.

Percentile Based Scores
☐ Centiles
☐ 5-grade classification: 10:20:40:20:10 centile splits
☐ Deciles
☐ Other *(describe)*: Click or tap here to enter text.

Standard Scores
☐ College Entrance Examination Board (e.g. SAT mean=500, SD=100)
☐ C-scores
☐ IQ deviation quotients etc. (e.g. mean 100, SD=15 for Wechsler or 16 for Stanford-Binet)
☐ Stanines
☐ Stens
☐ T-scores
☐ Z-scores

Other
☐ Critical scores, expectancy tables or other specific decision-oriented indices
☐ Raw score use only
☐ Other *(describe)*: Click or tap here to enter text.

3.4

Score transformation for standard scores
Scores are normalised when a non-linear transformation is applied to make a previously non-normal distribution, normal. In practice this usually means that a look up table is required to convert a score to the standard scale. When scores are not-normalised a simple linear transformation can be applied without a look up table, although in practice a look up table may be used in this situation as well.

☐ Standard scores obtained by linear transformation
☐ Standard scores obtained by use of normalisation look-up table
☐ Not applicable

3.5	**Continuous norming procedures**	☐ Age specific norms not used ☐ Age specific norms provided from separate age-delineated samples ☐ Continuous norming used to provide age specific norms

4. Digitally-generated reports

4.1 **Are digitally generated reports available with the instrument?**

If there is more than one report, please complete a separate form for each report.

☐ Yes *(complete items below)*
☐ No *(move to item 5.1)*
☐ If Yes, how many different reports are available? Click or tap here to enter text.

4.2 **Name or description of report**

Click or tap here to enter text.

4.3 **Design or presentation**

Select all that apply.

☐ Graphics only ☐ Integrated text and graphics ☐ Sound version available ☐ Text only
☐ Unrelated text and graphics ☐ Video version available

4.4 **Structure**

Select all that apply.

Some reports generate a text unit for a sten score in a scale-by-scale description. Others generate text units which relate to patterns or configurations of scale scores and consider scale interaction effects.

☐ Construct-based: built around one or more sets of constructs (typology) derived from original/base scale scores.
☐ Criterion-based: where the report focuses on links to empirical outcomes.
☐ Factor-based: where the report is constructed around higher order factors such as the Big Five in a personality measure.
☐ Pattern-based: e.g. descriptions of patterns and configurations of scale scores, and scale interactions
☐ Scale-based: e.g. a list of paragraphs giving scale descriptions
☐ Other *(describe)*: Click or tap here to enter text.

4.5 **Sensitivity to context**

Select one.

Reports generated from the same test but for different audiences, different purposes in different areas of activity will use different language, information, and design.

☐ One version for all contexts
☐ User definable contexts
☐ Pre-defined context-related versions *(list available contexts)*: Click or tap here to enter text.

4.6 **Development of the report**

Select all that apply.

The content (text units, etc.) of some report systems is based on the judgment of one or

☐ AI generated empirical/actuarial relationships
☐ Based on expert analysis of empirical/actuarial relationships

	more people who are 'expert-users' of the instrument. Others link scale scores to, for example, job performance measures, clinical classification, etc. while the content of some is generated by digital techniques with no initially identifiable author.	☐ Based on expert judgment of group of experts ☐ Based on expert judgment of one expert
4.7	**Modifiability** Select one.	☐ Not modifiable ☐ Limited modification (limited to certain areas, e.g. biodata fields) ☐ Unlimited modification
4.8	**Transparency** Select one.	☐ Clear linkage between constructs, scores, and text ☐ Concealed link between constructs, scores, and text – a 'black box' ☐ Mixture of clear/concealed linkage between constructs, scores, and text
4.9	**Type of content** Select all that apply. ☐ Not explicitly stated ☐ Behavioural descriptions ☐ Competence descriptions ☐ Diagnostic categories ☐ Questions for consideration ☐ Suggested future actions ☐ Suggested discussion points ☐ Other *(describe)*: Click or tap here to enter text.	
4.10	**Intended recipients** Select all that apply. *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). ☐ Qualified system users ☐ Qualified test users ☐ Test takers ☐ Third parties ☐ Other *(describe):* Click or tap here to enter text.	

5. Supply arrangements and materials

Information given in this section is likely to go out of date quickly. Publishers change rapidly, update and add to tests frequently and prices change regularly. 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.

5.1	**Supporting information provided by the distributor to users** Select all that apply.	□ Books and articles of related interest □ Discussion/User Groups □ Supplementary technical information and updates (e.g. local norms, local validation studies etc.) □ Technical manual □ User Manual □ Other *(describe)*: Click or tap here to enter text.
5.2	**Methods of publication** Select all that apply. □ Website □ Downloadable documents □ Print documents □ Other *(describe)*: Click or tap here to enter text.	
5.5	**Test-related qualifications required by the supplier of the test** Describe the requirements of the publisher. Examples might be: - EFPA or other European related qualifications - Practitioner psychologist - Specific national or professional accreditation - Test specific accreditation Where qualification requirements are not clear this should be stated. When it is explicitly stated that there are no qualification, write '*none*'. Click or tap here to enter text.	

Short stand-alone non-evaluative description

A concise non-evaluative description of the instrument. This should provide the reader with a clear idea of what the instrument claims to be. It should be as objective and factual as possible in tone, describing what the instrument is, the scales it measures, its intended purpose, the availability and type of norm groups, general points of interest or unusual features and any relevant historical background. It should also indicate who the intended test users and takers are. This description may be quite short (200-300 words). However, for more complex multi-scale instruments, it will need to be longer (300-600 words in most cases but it may need to be longer for tests with many versions and reports). It should be written so that it can stand alone as a description of the instrument in other contexts. As a consequence it may repeat some of the more specific information provided in response to sections 1-5. It should outline all versions of the instrument that are available and referred to on subsequent pages.

Click or tap here to enter text.

Part 2. Evaluation of the instrument

Information sources

Information sources that might inform these reviews, include:

- Manuals, white papers, website material, sample questions and reports that are supplied by the publisher for the user. They form core resources for the review.
- Open information that is available in academic or other literature, such as journal articles and books on testing, whether in printed or other formats: the reviewer usually sources this and may make use of this information in the review.
- Information held by the distributor/publisher that is not formally published or made available. The distributor/publisher may offer this at the outset or supply it when the review is sent 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.
- Information that is commercially confidential. In some instances, publishers may have technically important material that they are unwilling to make public for commercial as well as copyright and intellectual property reasons. 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 should enter 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.” In such situations, the reviewers’ non-competitive position against the test publisher becomes critical.

Explanation of ratings

All sections (unless other wise indicated) are scored using the following rating system. 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 for the stated purpose, 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 for the purpose it is intended to fulfil. This does not indicate any more general evaluation.

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**

Note. Users can combine 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. 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.

6. Quality of the explanation of the rationale, its 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. The term 'documentation' is taken to cover all those materials supplied or readily available to the qualified user: e.g. the manual; technical handbooks; booklets of norms; manual supplements; updates from publishers/suppliers and so on.

6.1. Rationale and development

This section relates to the procedures followed in the development of the instrument including developing a rationale, appropriate content for its measurement and adequate and appropriate analysis at the granular level of tasks or items. It is not always easy to rate these aspects of a test before looking at other aspects in the review. It may be worth examining this section after completing other sections.

Items to be rated n/a or 0 to 4		Rating					
6.1.1	**Theoretical foundations of the constructs**	n/a	0	1	2	3	4
6.1.2	**Summary of empirical research relating to the construct**	n/a	0	1	2	3	4
6.1.3	**Test development procedure** This includes both qualitative procedures as well as quantitative analyses.	n/a	0	1	2	3	4
6.1.4	**Translation or adaptation procedure (see e.g.** Iliescu, 2017)	n/a	0	1	2	3	4
6.1.5	**Thoroughness of analyses on test content** (including item analysis model and item analyses).						
	Not applicable	n/a					
	No information given	0					
	Insufficient item analysis carried out or results poor; e.g. discrimination low for many items; ceiling effects on scores	1					
	Basic item analysis procedure carried out and results adequate (item discrimination high enough given length of test)	2					
	Detailed item analysis procedure carried out and results adequate or better (item discrimination high enough given length of test)	3					

	Detailed item analysis procedure carried out, all items discriminate well and support score variance. For multi-scale instruments absence of unexpected cross-loadings.	4					
6.1.6	**Procedures to develop item content including considerations of content validity**	n/a	0	1	2	3	4
6.1.7	**Overall rating of the quality of the rationale and development** This overall rating is obtained by using judgment based on the ratings given for items 6.1.1 – 6.1.6.	n/a	0	1	2	3	4

6.2. Adequacy of documentation available to the user

This section covers the comprehensiveness and clarity of the coverage and explanation of the documentation available to the user (user and technical manuals, norm supplements, etc.).

The quality of the instrument itself, as evidenced by the documentation, is treated in sections : 6.1, 6.3, 6.4, 9, 10 and 11. The reviewer may want to complete those sections first and then return to complete this section 6.2.

	‘Benchmarks’ are provided for an ‘excellent’ (4) rating.	Rating					
6.2.1	**Rationale** *(see rating 6.1.7)* 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
6.2.2	**Development** Excellent: Full details are given of item sources, development of stimulus material according to accepted guidelines, piloting, item analyses, comparison studies and changes made during development trials.	n/a	0	1	2	3	4
6.2.2	**Development of the test through translation/adaptation** Excellent: Information in the manual shows that the translation/adaptation process was done according to international guidelines (ITC, 2017) and included: input from native speakers of new language; multiple review by both language and content (of test) experts; independent checks of quality of translation/adaptation; consideration of cultural and linguistic differences.	n/a	0	1	2	3	4

6.2.3	**Standardisation** Excellent: Clear and detailed information is provided about sizes and sources of standardisation sample and standardisation procedure.	n/a	0	1	2	3	4
6.2.4	**Norms** Excellent: Clear and detailed information is provided about sizes and sources of norms groups, representativeness, conditions of assessment, any algorithms, or procedures in the norming process that impact interpretation.	n/a	0	1	2	3	4
6.2.5	**Reliability / Precision** Excellent: Clear and detailed explanation of how reliability / precision was assessed, results of analyses and the appropriateness of the approach(es) used given the nature of the instrument.	n/a	0	1	2	3	4
6.2.6	**Validity based on internal structure.** Excellent: Gives clear and detailed explanation of validity based on internal structure with a wide range of studies clearly and fairly described.	n/a	0	1	2	3	4
6.2.7	**(a) Validity based on relations with other variables** Excellent: Gives clear and detailed explanation of validity based on relations with other variables, with a wide range of studies clearly and fairly described.	n/a	0	1	2	3	4
	(b) Validity based on other sources Excellent: Gives clear and detailed explanation of other sources of validity with a wide range of studies clearly and fairly described.	n/a	0	1	2	3	4
6.2.8	**Digitally generated reports** Excellent: Gives clear and detailed information about the format, scope, reliability, and validity of computer-generated reports. This should also cover the language used and whether it is inclusive for diverse stakeholders.	n/a	0	1	2	3	4
6.2.9	**Language** Uses inclusive, non-discriminatory language throughout.	n/a	0	1	2	3	4
6.2.10	**Adequacy of documentation available to the user** This rating is obtained by using judgment based on the ratings given for items 6.2.1 – 6.2.9.	n/a	0	1	2	3	4

6.3. Quality of the procedural instructions provided for the user

‘Benchmarks’ are provided for an ‘excellent’ (4) rating		Rating					
6.3.1	**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
6.3.2	**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
6.3.3	**Norming** Excellent: Clear and detailed information provided, with checks described to deal with possible wrong norm groups and errors in score transformations. If transformation of raw scores into standard scores is done automatically, there is evidence that score transformation is correct and the right norm group is applied.	n/a	0	1	2	3	4
6.3.4	**Interpretation and reporting** Excellent: Detailed advice is provided 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
6.3.5	**Providing feedback and debriefing test takers and others** Excellent: Detailed advice provided on how to present feedback to candidates including the use of computer-generated reports if available.	n/a	0	1	2	3	4
6.3.6	**Providing good practice issues on fairness and bias** Excellent: Detailed information is provided about work done to assess the existence of any bias with respect to different groups and if found, work done to address this and implications for use.	n/a	0	1	2	3	4

6.3.7	**Restrictions on use** Excellent: Clear descriptions given of who should and who should not be assessed, with well-explained justifications for restrictions, for instance literacy levels required.	n/a	0	1	2	3	4
6.3.8	**Software and technical support** Excellent: In the case of technology assisted testing, there is a clear description of software and hardware requirements, the operation of the software (covering possible errors and use of different systems), and availability of technical support.	n/a	0	1	2	3	4
6.3.9	**References and supporting materials** Excellent: Detailed references provided to the relevant supporting academic literature and cross-references to other related assessment instrument materials.	n/a	0	1	2	3	4
6.3.10	**Quality of the procedural instructions provided for the user** This overall rating is obtained by using judgment based on the ratings given for items 6.3.1 – 6.3.9.	n/a	0	1	2	3	4

6.4. Overall adequacy

6.4	This overall rating is obtained by using judgment based on the ratings given for the sub-sections 6.1, 6.2, and 6.3.	
	Not applicable	n/a
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

7. Quality of the test materials

7.1. Quality of technology-enabled test materials

This sub-section can be **skipped** if not applicable.

Items to be rated n/a or 0 to 4		Rating					
7.1.1	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
7.1.2	Ease with which the test taker can understand the task.	n/a	0	1	2	3	4
7.1.3	Clarity and comprehensiveness of the instructions (including sample items and practice trials) for the test taker, the operation of any software, and how to respond.	n/a	0	1	2	3	4
7.1.4	Ease with which responses or answers can be made by the test taker.	n/a	0	1	2	3	4
7.1.5	Quality of the design of the user interface.	n/a	0	1	2	3	4
7.1.6	Accessibility of the test for differently-abled test takers. Excellent: 'Accessible by design' approach used in developing materials and with broad accessibility of the test.	n/a	0	1	2	3	4
7.1.7	Quality of item content (use of language, quality of graphics, or objects used in the test).	n/a	0	1	2	3	4
7.1.8	Quality of technology enabled materials This overall rating is obtained by using judgment based on the ratings given for items 7.1.1 – 7.1.7.	n/a	0	1	2	3	4

7.2. Quality of paper-&-pencil and other non-technology enabled test materials

This sub-section can be **skipped** if not applicable.

Items to be rated n/a or 0 to 4		Rating					
7.2.1	Quality 'look and feel' of the test materials (test booklets, answer sheets, test objects, etc.).	n/a	0	1	2	3	4

7.2.2	Ease with which the test taker can understand the task.	n/a	0	1	2	3	4
7.2.3	Clarity and comprehensiveness of the instructions (including sample items and practice trials) for the test taker.	n/a	0	1	2	3	4
7.2.4	Ease with which the test taker can understand the task.	n/a	0	1	2	3	4
7.2.5	Ease with which responses or answers can be made bv the test taker.	n/a	0	1	2	3	4
7.2.6	Accessibility of the test for differently-abled test takers. Excellent: 'Accessible by design' approach used in developing materials and with broad accessibility of the test.	n/a	0	1	2	3	4
7.2.7	Quality of item content (use of language, quality of graphics or objects used in the test).	n/a	0	1	2	3	4
7.2.8	Quality of the materials for paper and pencil and other non-technology enabled tests. This overall rating is obtained by using judgment based on the ratings given for items 7.2.1 – 7.2.7.	n/a	0	1	2	3	4

Reviewers' comments on quality of the materials

Click or tap here to enter text.

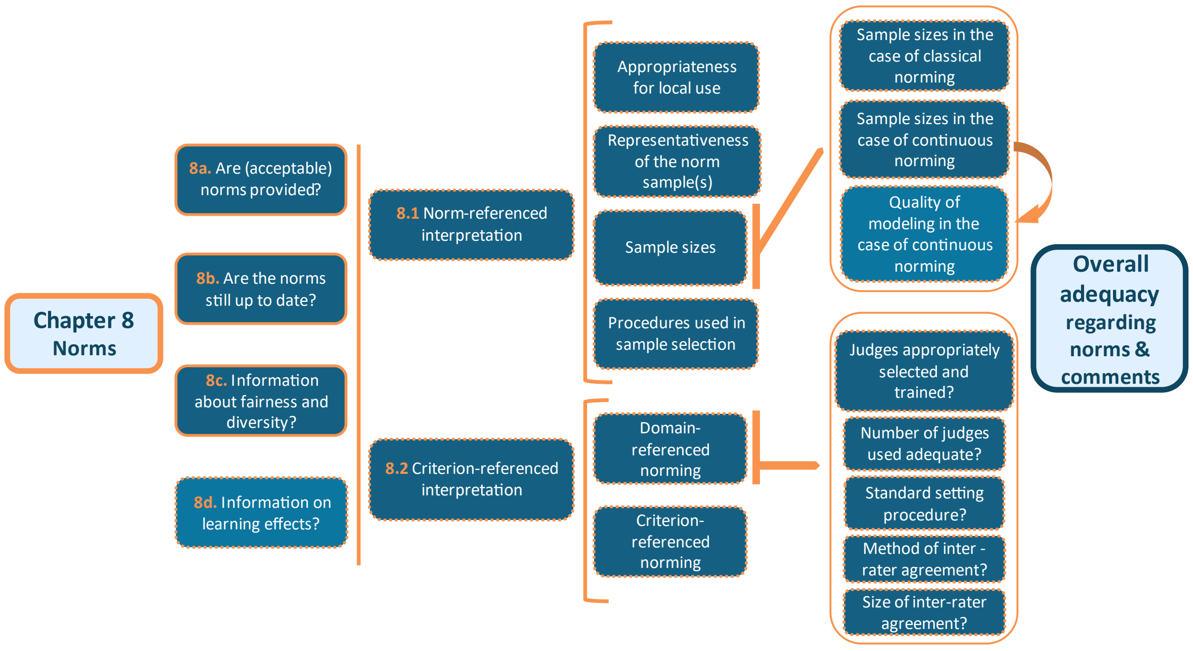
General guidance on assigning ratings for the next 3 chapters

It is difficult to set clear criteria for rating the more technical, psychometric qualities of an instrument including norms, reliability, and validity. The notes in chapter 8, 9 and 10 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 used, as well as the types of decisions to be made using the instrument, will all affect the way in which ratings are awarded.

Metaphorically, we can say that we have provided a recipe for a carbonara, for example, and distinguished between 'need to have' and 'nice to have' ingredients, but the cook, the kitchen, the equipment, the available time, the region etc. may play a role.

8. Norms

Introduction


Chapter 8 Norms
8a. Are (acceptable) norms provided?
8b. Are the norms still up to date?
8c. Information about fairness and diversity?
8d. Information on learning effects?
8.1 Norm-referenced interpretation
8.2 Criterion-referenced interpretation
Appropriateness for local use
Representativeness of the norm sample(s)
Sample sizes
Procedures used in sample selection
Domain-referenced norming
Criterion-referenced norming
Sample sizes in the case of classical norming
Sample sizes in the case of continuous norming
Quality of modeling in the case of continuous norming
Judges appropriately selected and trained?
Number of judges used adequate?
Standard setting procedure?
Method of inter - rater agreement?
Size of inter-rater agreement?
Overall adequacy regarding norms & comments

We can distinguish two ways of scaling or categorizing raw test scores (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

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 8.1).

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 8.2).

Raw scores will be categorized in two or more different score ranges in the two latter ways of working, e.g., ‘pass’ or ‘fail;’ assigning patients in different score ranges to different treatment programs; assigning pupils scoring below a critical score to remedial teaching; or accepting or rejecting applicants in personnel selection. If multiple (kinds of) norms are provided, the ratings for different norm groups can be repeated. **In all cases** the next three criteria described below (8a, 8b and 8c) will have direct influence (‘***need to have’***) on the final rating of the norms.

(8a) Are (acceptable) norms provided?

To assess the quality of the norms provided, an adequate and complete description of the method of sampling or data collection must be provided. An adequate and complete description of the data collection method consists at least of a clear description of the procedure followed, the response rate, the collection situation and, if applicable, the collection period and whether the collection was 'unproctored' or supervised.

Obviously, norms should be available for the test when the test is marketed and should be suitable for use.

The following situations may lead to examples of lack of norms or inadequate norms. Multiple of these cases listed below are inspired by examples given in the (new, in press) COTAN Review System (Egberink & Leng, 2009-2024).

- Norm tables or cut-off scores are not provided. Information about the score distribution (e.g., averages, standard deviations, skewness, kurtosis) alone is not adequate, as it does not provide sufficient guidance for the user to interpret every possible raw score easily and without error.
- After the norm data were collected, substantial changes were made to the test itself, for example changes to the items or instructions.
- The norm data were collected using a different test mode or version, and research on the equivalence of the two versions is lacking or insufficient (Bugbee Jr, 2014). For example, paper-and-pencil data collection where, in use, the test will be administered digitally (or vice versa) will not be acceptable without evidence of equivalence. It cannot be assumed that ability and skill tests and/or tests bound by a time limit will be equivalent and new norm data should be collected or equivalence plausibly demonstrated. For non-ability tests such as personality questionnaires, the mode of administration appears to have minor influence on the value of norms so less evidence of equivalence is needed (Bartram, 2005; King & Miles, 1995; Mead & Drasgow, 1993).
- The conditions in which the norm data were collected differ substantially from the conditions in which the test will be administered, e.g. low - versus high-stakes conditions; with or without a time limit; or proctored versus unproctored administration conditions.
- In tests intended for group-level interpretation, norm tables are provided based on individual scores, or vice versa.
- The norm data were not collected from the reference population appropriate for the test's purpose.
- Judgments of appropriateness for intended applications should be clearly described. This should include whether the type of norm scale is consistent with the purpose of the test and whether the interpretation and limitations of the norm scale are made clear to the user.

Are (acceptable) norms provided? (please comment)
Click or tap here to enter text.

(8b) Are the norms still up to date?

Both relative and absolute norms are subject to wear and tear. Of the psychometric characteristics of a test, norms are the most sensitive to societal changes (see, for example, the Flynn effect; Trahan, Stuebing, Fletcher, & Hiscock, 2014), changes in education, change in DSM criteria and in job content.

Therefore, renorming of the test should take place from time to time, or the author should demonstrate through research that renorming is not necessary. However, there is a lack of consensus regarding how long norms remain valid. The German assessment system for test quality (Hagemeister, Kersting & Stemmler, 2012) recommends an eight-year period for renorming, incidentally without attaching any consequences. In the Spanish test review model there is an additional rating point 'Adequate with shortcomings' for norms of 20 to 24 years old and finds '+25 years' unacceptable. This suggests up to 20 years before renorming is reasonable.

The APA Standards (APA,2014, p. 104, Standard 5.11) state that: "*... so long as the test remains in print, it is the publisher's responsibility to renorm the test with sufficient frequency to permit continued accurate and appropriate score interpretations*". The APA does not specify a deadline in this regard.

Balancing between what is practically feasible and desirable, the Dutch COTAN alerts users to potentially worn-out standards with the footnote "The standards are obsolete" to the assessment of the norms of tests for which re-norming or calibration research has not taken place for 15 years. After another five years without such examination, the footnote is changed to: "*Due to obsolescence, the norms are no longer usable*," and the rating becomes "*unsatisfactory*"."

To allow the reviewer to assess whether norms are current or potentially outdated, it is important to mention the year or period of data collection. To assess timeliness/obsolescence, the year in which most of the data was collected is taken as the starting point. The absence of information about when the data were collected, should result in a rating of '*insufficient information*'.

If a re-norming study has been carried out, it is expected that as well as providing revised norms, the new norm data will be used to confirm relevant indicators of Reliability and Validity (i.e. internal structure).

The EFPA Test Review Model uses the following guidelines for ratings related to norms, however reviewers should take into account the context of use of the test and adjust accordingly. For example, an original standardization norm with a minimal sample size for a well-used test might be expected to be revised earlier than a norm for an older test which has been subject to years of studies showing little if any movement in the norms over the years.

How old are the normative studies?	
[n/a]	Not applicable to this instrument
[0]	No (or insufficient) information is provided
[1]	Inadequate: 20 years or older
[2]	Adequate: norms between 15 and 19 years old
[3]	Good: norms between 10 and 14 years old
[4]	Excellent: norms less than 10 years old

(8c) Information about fairness and diversity

It is important that attention is paid, and information provided about minority/protected group performance comparisons. The groups studied will depend on the nature of the instrument. Effects of age and gender will usually be expected but there are many other relevant variables including, nationality, language, ethnic identification, disability (various), socio-economic status, educational opportunity etc. The review should consider which groups are relevant considering the nature of the reference population and the use of the instrument. The following is a guide to ratings:

Information about fairness and diversity	
[n/a]	Not applicable to this instrument
[0]	No information is provided
[1]	Inadequate information is provided
[2]	Adequate: general information is provided, with minimal information/analysis
[3]	Good: analyses of data for most relevant groups with clear descriptions of findings including differences where found
[4]	Excellent: good range of analyses with clear and fair discussion of results and relevant issues relating to use and interpretation

It is desirable, but not required ('nice to have'), that the following issue(s) are considered in developing and presenting norm data and advice on test use.

(8d) Information on learning effects for the instrument and implications for retesting

How old are the normative studies?	
[n/a]	Not applicable to this instrument

[0]	No information is provided
[1]	Inadequate information is provided
[2]	Adequate: general advice about practice effects but no specific information for the instrument
[3]	Good: some test specific information given e.g., regarding time lapse advised for subsequent testing
[4]	Excellent: quantified information provided about impact of multiple test administration on scores. Norms for second test application after typical test-retest-interval provided

8.1. Norm-referenced interpretation

This sub-section can be **skipped** if not applicable.

Criteria about the quality of the norms in the case of norm-referenced interpretation

8.1.1	**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 to this instrument	[n/a]
No information is provided	[0]
Not locally relevant (e.g., inappropriate foreign or international samples)	[1]
Sample(s) that do(es) not fit well with the relevant application domain but could be used with caution (may include international samples where it is reasonable to assume only minor impact of language or culture)	[2]
Local country samples (or relevant international samples where international comparison is required) with good relevance for intended application	[3]
Local country samples (or relevant international samples where international comparison is required) drawn from well-defined populations from the relevant application domain	[4]

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 chapter 8.

An international norm may be the most appropriate for international usage with different languages test versions, 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 subset 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 of translation. This will apply to performance tests and to some extent to abstract and diagrammatic reasoning tests.

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.
- There should be evidence of measurement invariance across language groups.
- Information should be provided regarding whether 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 needs 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 including consideration of whether international norm groups are based on all individuals completing the same language version or different language versions.

Representativeness of the norm sample(s)

A norm group must be representative of the reference 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, ethnicity, purpose and conditions of testing) is similar to that of the population. The use of an appropriate probability sampling model will tend to enhance representativeness.

8.1.2

It may happen that the distribution of a variable in the sample does not match that in the population. There are a number of ways to deal with this.

By showing (if it is the case) that the test scores do not differ meaningfully across the different levels of any relevant variable, in statistical tests with sufficient power.

If differences do exist, then it may be possible to address imbalances in representativeness through *weighting*. *Weighting* involves using a weight for each person in the calculation of the norms, such that in the weighted sample, the distributions of the relevant variables are more similar to those of the population. In the case of under-representation of a subgroup, a weight greater than 1 may be given to each test taker and, in the case of over-representation, a weight less than 1. In the case of under-representation, the weighting factor should not exceed 2. In case of over-representation, weighting is preferable done by randomly removing individuals from the sample.

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

8.1.3 **Sample sizes**

8.1.3.1 Sample sizes in the case of 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 these may be the score ranges that are most relevant for decisions to be taken). 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).

The sample size requirement applies to each norm group.

	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

8.1.3.2 Sample sizes in the case of continuous norming

Continuous norming procedures are increasingly being applied. They are used particularly for tests that are intended for use in schools 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 level of accuracy for multiple norm groups. Bechger, Hemker, and Maris (2009) have computed some values for the sizes of continuous norm groups that would give equivalent accuracy compared to classical norming. When eight sub-groups are used N = 70 (8x70) gives the same accuracy as N =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 using 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 some strict statistical assumptions. The test author has at least 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 one step higher.

More information on sampling issues when continuous norming is applied, to be found via: Innocenti, Tan, Candel, & Van Breukelen (2023) and Timmerman, Voncken, & Albers (2021).

	Low-stakes use	High-stakes decisions	
Not applicable			n/a
No information given			0
Inadequate sample size	e.g., fewer than 8 subgroups with a maximum of 69 respondents each	e.g., 8 subgroups with 70 - 99 respondents each	1
Adequate sample size	e.g., 8 subgroups with 70 - 99 respondents each	e.g., 8 subgroups with 100 - 149 respondents	2

Good sample size	e.g., 8 subgroups with 100 - 149 respondents	e.g., 8 subgroups with at least 150 respondents each	3
Excellent sample size	e.g., 8 subgroups with at least 150 respondents each		4

8.1.3.3 Quality of Modeling in the case of continuous norming

If continuous norming is applied, the quality of the norms depends on the suitability of the applied continuous norming model, in addition to the size and representativeness of the norm data. Therefore, the evaluation of the norming model and how well its assumptions are met, must be considered in rating the norms. Some guidance is provided here but for more background on continuous norming, please refer to relevant literature (e.g., Bechger, Hemker, & Maris, 2009; Oosterhuis, Van der Ark, & Sijtsma, 2016; Timmerman, Voncken, & Albers, 2021).

Based on issues regarding the quality of modeling in this case formulated in the (new, in press) COTAN Review System (Egberink & Leng, 2009-2024), the following four considerations are important in evaluating the applied continuous norming model.

1. The norming model used should be clearly described. The functional form (e.g., linear, higher-order polynomial) of the relationship between the reference predictor(s) used and the parameters of the test score distribution to be modelled should be specified. The selection procedure that was used to obtain the final norming model, should be described. This should include an explanation of how well the model selected fits the empirical data and how the risk of overfitting was addressed in selecting the chosen model.
    a. Model comparison tests should show which reference predictor(s) (including which powers and predictor products) were (initially) included in the different models and the selection criteria (e.g., fit diagnostics like the Akaike Information Criterion- AIC; cross-validation) that were used to determine the final model.
    b. Visual inspection procedures (e.g., centile curve plots, worm plots) should be presented and explained.
    c. Theoretical considerations and assumptions should be clearly explained (e.g., the expectation that the relationship between the reference predictor(s) and the test score is monotonically increasing).
2. In some cases, after applying a continuous norming model, additional corrections are made, for instance, to match the estimated function to theoretical expectations or so-called smoothing. It should be clearly reported if and how any corrections or smoothing was applied, and what the implications are with respect to the modelled parameters.
3. The fit of the selected norming model should be described. The fit of the final model can be demonstrated using statistical fit measures and can be represented graphically (e.g., using worm plots or other plots where the values implied by the final model are plotted against the observed values). The reported results should be clearly interpreted and explained. Because of the potential complexity and flexibility of continuous norming models, it is important to explain how overfitting (i.e., the model is too focused on a specific sample and therefore cannot be generalized well to other samples from the target population) has been avoided.

4. The conversion of conditional score distributions into norm scores should be described. Here, it is important that the choice and unit of the reference predictor(s) (e.g., in the case of 'age', day/week/month/other) is justified.

Finally, we are in agreement with the COTAN (Egberink & Leng, 2009-2024), that the use of continuous norming to extrapolate beyond the observed range of the reference predictor(s) is not appropriate. Such an application should be assessed as 'inadequate'.

Not applicable	n/a
No information given	0
Inadequate: Modelling procedures are not appropriate or not well implemented or extrapolation beyond the observed data range	1
Adequate: Appropriate modelling procedures but little detail of process or inadequate checks for fit etc.	2
Good: Appropriate modelling procedures and model checks but some details missing	3
Excellent: Appropriate modelling procedures and good description showing how model suitability and fit was checked, and norms created	4

8.1.4 **Procedures used in sample selection** *(select one or more)*

When the sample is gathered with a probability sampling model the chance of being included in the sample is equal for each element in the population and this is the preferred sampling choice. 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 also multiple methods can be differentiated: pure convenience sampling (this is often simply adding every tested person to the norm group, as is done in many samples for personnel selection; post-hoc data may be classified into meaningful sub-groups based on biographical and situational information); dynamic sampling (a convenience sample of every person tested where the norm is continuously updated as more data is collected); quota sampling (as in convenience sampling, but the proportion of respondents in each subgroup required is specified in advance, similar to survey research procedures); snowball sampling (asking contacts to participate, who in turn approach their contacts, etc.) and purposive sampling (e.g., selecting particular diagnostic groups to participate). In all these cases of non-probability sampling the rationale for this choice and the consequences for representativeness of the norm group(s) must be closely monitored and reviewed.

The appropriateness of sample selection procedures should be commented upon in the Reviewers Comments at the end of the section

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 – convenience with dynamic updating	[ ]
Non-probability sample – quota	[ ]
Non-probability sample – ‘snowball’	[ ]
Non-probability sample – purposive	[ ]

	Other *(describe)*: Click or tap here to enter text.	[ ]

8.2. 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 8.2.1) 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 8.2.2).

8.2.1 Domain-referenced norming

8.2.1.1 **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

8.2.1.2 **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 task, the context, and the level of expertise of the judges. The suggested number of judges is for typically scenarios but what is appropriate will vary.

Not applicable	n/a
No information given	0
Inadequate (typically less than 5 judges)	1
Adequate (e.g., 5-9 judges)	2
Good (e.g., 10-14 judges)	3
Excellent (e.g., 15 or more)	4

8.2.1.3	**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)*: Click or tap here to enter text.	[ ]
8.2.1.4	**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)*: Click or tap here to enter text.	[ ]
8.2.1.5	**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 ≤ r < 0.70)	2
	Good (e.g., 0.70 ≤ r < 0.80)	3
	Excellent (e.g., r ≥ 0.80)	4

8.2.2	Criterion-referenced norming	
	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 the judgment. 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 should also be taken into account.	
	Not applicable	n/a
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

8.3. Overall adequacy

8.3	This overall rating is obtained by using judgment based on the ratings given for items 8.1 to 8.2.2	
	Not applicable	n/a
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

The overall rating for norm-referenced interpretation can in most cases be no higher than the rating for 8.1.1, 8.1.2 and 8.1.3, but it can be lower dependent on the other information provided.

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 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 8.2.2, but it can be lower e.g., when the studies are too old.

Reviewers' comments on the norms
Brief report about the norms and 'their history,' including e.g., informa-tion on provisions made by the publisher/author for updating norms on a regular basis. Underline the strong and weak aspects of the quality of the norms.

Click or tap here to enter text.

9. Reliability/Precision

General guidance on assigning ratings for this section

Reliability/precision refers to the degree to which scores are free from measurement error variance. In other words, reliability/precision describe consistency of test scores. For reliability/precision, the guidelines are based on the need to have a small Standard Error for estimates of it. Guideline criteria for reliability/precision are given in relation to two distinct contexts: the use of instruments to make decisions about groups of people and for individual assessments. Reliability/precision requirements are higher for the latter than the former. Other factors can also affect reliability/precision 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 coefficients of the composite should be the focus for rating, not the reliabilities of the components.

For some exercises, such as game-based assessments, standard reliability models may be difficult to apply. There may be thousands of data points collected and/or the dynamic nature of the task may mean individual test taker experiences are very different. However, this does not remove the need to show reliability or precision of measurement – and indication of the standard error of the score or that the same individual tested on two occasions, or two individuals with the same level of the construct measured would receive the similar scores. While item-based indicators such as Omega or Cronbach's alpha or IRT information measures may not be suitable, parallel form, test-retest or even split half paradigms could be appropriate. If none of these methods suit, the test authors need to provide sufficient alternative evidence of the reliability or precision of scores to allow its use (Burstein, 2023)[1]. The test review will need to evaluate whether the evidence provided is sufficient and provide ratings in section 9.7 with additional comments on the suitability of the approach.

When an instrument has been translated and/or adapted from a non-local context, one could apply the original version's reliability/precision evidence 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. However, for internal consistency, reliability evidence based on local groups is preferable, 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 can be found in The ITC Guidelines for Translating and Adapting Tests (Second edition) (International Test Commission 2017).

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 an instrument, its area of application, the quality of the data on which reliability/precision 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 coefficient 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

[1] The Duolingo English Test Responsible AI Standards. [Updated March 29, 2024]. Go to https://go.duolingo.com/ResponsibleAI.

averaging sets of ratings (there is space for this at the end of the chapter summarizing the reliability estimates).

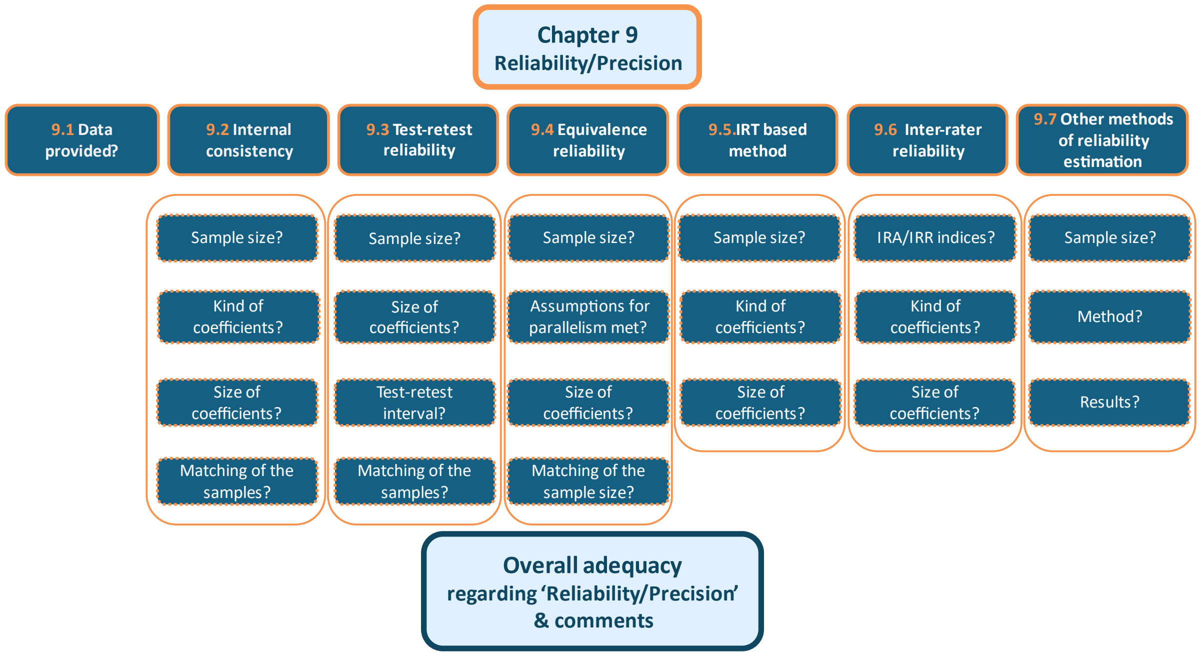
In order to provide some idea of the range and distribution of values associated with the various scales that make up an instrument, enter the number of scales in each section. For example, if an instrument being used for group-level decisions had 15 scales of which five had retest reliabilities coefficients 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 ≤ r < 0.70)	[6]	2
Good (e.g. 0.70 ≤ r < 0.80)	[4]	3
Excellent (e.g. r ≥ 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 coefficients than their component primary scales. As an example, primary scales in a multi-scale instrument may have reliability coefficients around 0.70 while Big Five secondary aggregate scales based on these can have reliability coefficients 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 at the end of this chapter. 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.’



9.1 Data provided about reliability/precision *(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	[ ]

9.2 Internal consistency

The use of internal consistency coefficients is not sensible for assessing the reliability/precision 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/precision of ipsative scales. Alternate form or retest measures are more appropriate for these scale types.

Internal consistency coefficients give a better estimate of reliability/precision 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 9.7.

9.2.1	**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 adequately sized study		3
	Good range of adequate to large studies		4
9.2.2	**Kind of coefficients reported** *(select all that apply)*		
	Not applicable		n/a
	Coefficient alpha or KR-20		[ ]
	Lambda-2		[ ]
	Greatest lower bound		[ ]
	Omega (factor analysis)		[ ]
	Theta (factor analysis)		[ ]
	Other, describe: Click or tap here to enter text.		[ ]
9.2.3	**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

9.2.4	**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	[ ]
	No information given	[ ]
	Not applicable	n/a

9.3 **Test-retest reliability/precision – 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/precision of state. In this case all items concerning test-retest reliability/precision should be marked '*not applicable.*'

9.3.1	**Sample size**	
	Not applicable	n/a
	No information given	0
	One inadequate study (typically sample size less than 100)	1
	One adequate study (typically sample size of 100-200)	2
	One large (typically sample size more than 200) or more than one adequately sized study	3
	Good range of adequate to large studies	4

9.3.2	**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. $0.60 \le r < 0.70$)	[ ]	2
	Good (e.g. $0.70 \le r < 0.80$)	[ ]	3
	Excellent (e.g. $r \ge 0.80$)	[ ]	4
9.3.3	**Data provided about the test-retest interval** (*select or fill in test-retest interval*)		
	Not applicable		n/a
	No information given		[ ]
	The interval is: Click or tap here to enter text.		[ ]
9.3.4	**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		[ ]
	.... do not match the intended test takers, leading to less favourable coefficients (e.g. reduction by restriction of range)		[ ]
	.... match the intended test takers		[ ]
	No information given		[ ]
	Not applicable		n/a
9.4	**Equivalence reliability/precision (parallel or alternate forms)**		
9.4.1	**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 adequately sized study		3

	Good range of adequate to large studies		4
9.4.2	**Are the assumptions for parallelism* met for the different versions of the test for which equivalence reliability/precision 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
9.4.3	**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 ≤ *r* < 0.80)	[ ]	2
	Good (e.g. 0.80 ≤ *r* < 0.90)	[ ]	3
	Excellent (e.g. *r* ≥ 0.90)	[ ]	4
9.4.4	**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		[ ]

	No information given	□
	Not applicable	n/a

9.5 IRT based method

9.5.1 **Sample size**

It is difficult to give uniform guidelines for the adequacy of sample sizes in case IRT methods for the estimation of reliability/precision 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 and the number of items, 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 adequately sized study	3
Good range of adequate to large studies	4

9.5.2 **Kind of coefficients reported** *(select all that apply)*

The first method gives the reliability/precision 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/precision 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: Click or tap here to enter text.	[ ]
Not applicable	n/a

9.5.3	**Size of coefficients based on the final test length**	**Number of scales** *(if applicable)*	**M***

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.

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 ≤ information < 5.00)	[ ]	2
Good (e.g. $0.80 \leq r < 0.90$; 5.00 ≤ information < 10.00)	[ ]	3
Excellent (e.g. $r \geq 0.90$; information ≥ 10.00)	[ ]	4

9.6 **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).

9.6.1	Note that it is important to distinguish between interrater agreement and interrater reliability. IRA/IRR indices should be selected in accordance with purpose and specificity of the analysis (Gisev, Bell, Chen, 2013, Kottner et al., 2011)	
	Not applicable	n/a
	No information given	0
	One inadequate study	1
	One study with small number of raters, sample size and correctly assessed level of measurement in terms of applied IRR/IRA index together with full explanation of the purpose of the analysis	2
	One study with large number of raters, sample size and correctly assessed level of measurement in terms of applied IRR/IRA index together with full explanation of the purpose of the analysis	3
	Good range of adequate to large studies using different ways to proof the degrees of agreement and/or inter-reliability	4
9.6.2	**Kind of coefficients reported** *(select all that apply)*	
	Not applicable	n/a
	Percentage agrees	[ ]
	Coefficient Kappa	[ ]
	Intra Class Correlation	[ ]
	Coefficient Iota	[ ]
	Other, describe: Click or tap here to enter text.	[ ]

9.6.3	**Size of coefficients** To some methods mentioned in 9.6.2 the guide numbers may not apply as no *r*'s are computed. Note that 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. Sometimes for Kappa's values more liberal standards are acceptable (e.g. see classification of Landis and Koch (1977). IRA/IRR values should be reviewed with consideration of the nature of the given study.	**Number of scales** *(if applicable)*	
	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
9.7	**Other methods of reliability or precision estimation** Required where item-based reliability/precision analysis is not suitable e.g. some game-based assessments. The reviewer will need to assess the methodology on its merits and the quality of the implementation.		
9.7.1	**Sample size**		
	Not applicable		n/a
	No information given		0
	One inadequate study		1
	One adequate study		2
	One large or more than one adequately sized study		3
	Good range of adequate to large studies		4
9.7.2	Describe method: Click or tap here to enter text.		

9.7.3	**Results**	**Number of scales** *(if applicable*)	**M***
	Not applicable		n/a
	No information given	[ ]	0
	Inadequate	[ ]	1
	Adequate	[ ]	2
	Good	[ ]	3
	Excellent	[ ]	4

9.8 **Overall Adequacy**

This overall rating is obtained by using judgment based on the ratings given for items 9.1 – 9.7.3. *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/precision 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/precision studies are reported
- whether also standard errors of measurement are provided
- procedural issues, e.g. group size, representativeness, number of reliability/precision studies, heterogeneity of the
- heterogeneity of group(s) on which the coefficient is 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/Precision Underline the strong and weak aspects of the evidence of reliability/precision available. Comments pertaining to equivalence/reliability/precision generalisation should also be made here (if applicable).
Click or tap here to enter text.

10. 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 classical 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. In fact, the most recent standards of the AERA, APA and NCME (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) highlight that it is not the test that is validated, but rather specific interpretations or uses of its scores. Thus, there are no different types of validity (content, construct, criterion, etc.) although different sources of validity evidence can be collected. The importance of collecting one or another type of evidence depends mainly on the intended use of the test. Borsboom et al. (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.

Whichever approach to validity one prefers, for a standardised judgment it is necessary to structure the concept of validity a bit. Of the various sources of evidence, we follow the terminology of AERA, APA and NCME (2014) and focus on the three most relevant types of evidence based on: (a) content; (b) relationships with other variables (with a criterion to be predicted, with another test measuring the same or a related construct, etc.); and (c) internal structure (e.g. assessing factor structure or the hypothesized dimensionality). The evidences may change 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 judgement is made about 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/developer/publisher should deliver. Note that the final ratings will be a kind of average of the provided evidence and that there may be situations or groups for which validity support may be stronger or weaker (or for which no evidence of validity has been provided at all). Notice that if the test manual uses the classical differentiation of different types of validity (e.g. content, construct, or criterion-referenced validity), the information should be incorporated into the appropriate section depending on the type of analysis performed.

For some digital exercises such as game-based assessments, particularly where a very empirical approach is taken to task design and what is measured, and content validity is low and/or evidence of internal structure is lacking, it is particularly important that strong validity evidence is provided to support the interpretation of scores. Patterns of relationships with relevant variables and criterion related validity are therefore even more important. While guidelines are provided for interpreting correlations between test scores and relevant criteria, reviewers can use their own judgement if results are expressed in terms of accuracy, precision, sensitivity (recall) and F1 scores. Comments should be provided regarding the appropriateness of the validity information given the intended use of the instrument.

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:

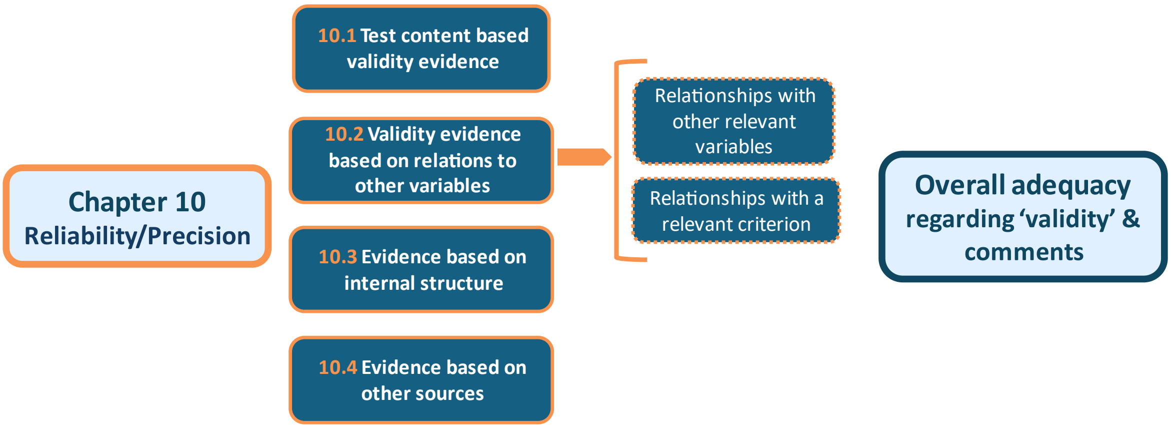
- Invariance in construct structure e.g. via factor structure or correlation with standard measures.
- Similar criterion related validity e.g. similar profile of correlations of a multi-scale instrument with in-dependent external criterion such as ratings of job competencies.
- Items show similar patterns 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 or metric invariance	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 10.1 and 10.2 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 valid-ity 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.


Chapter 10
Reliability/Precision
10.1 Test content based validity evidence
10.2 Validity evidence based on relations to other variables
10.3 Evidence based on internal structure
10.4 Evidence based on other sources
Relationships with other relevant variables
Relationships with a relevant criterion
Overall adequacy regarding 'validity' & comments

10.1. Test content based validity evidence

This aspect is especially essential in criterion-referenced tests and particularly in tests of academic performance. Make your judgement on the quality of the representation of the content or domain. If expert evaluations appear in the documentation provided, please take them into consideration. Data or arguments on content validity are treated in this assessment system as part of the test rationale and development process and have therefore already been addressed in section 6.

10.2. Validity evidence based on relations to other variables

It is crucial than the expected relationships, under the assumption that the test scores serve the intended purpose, are supported by the evidence based on the relationships with other variables. These may refer, for example, to different relevant groups, groups where there is a manipulation of the situation, other test scores, or a selected criterion that should be predicted by test scores. While criterion-related validity evidence alone may not be adequate to fully substantiate claims of validity for interpreting and using scores, this type of evidence can still be very valuable in constructing a comprehensive validation argument, depending on the test's purpose.

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, validity evidence based on the relationship between the test and a criterion 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), section 10.2.2. can be considered 'not applicable.'

10.2.1.	**Relationships with other relevant variables (instruments or groups)**	
10.2.1.1	**Designs and/or techniques employed** Select all that apply.	
	Not applicable	n/a
	No information given	[ ]
	Differences between groups	[ ]
	Correlations with other instruments	[ ]
	(Quasi-)Experimental Designs	[ ]
	Multi-Trait Multi-Method (MTMM) correlations	[ ]
	Other, describe:	[ ]
10.2.1.2	**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.	
	Not applicable	n/a
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
10.2.1.3	**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. When the expected differences are not shown, this would raise strong doubts about the valid use of the test to discriminate among relevant groups.	
	Not applicable	n/a
	No information given	0

	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
10.2.1.4	**Median and range of correlations between the test and tests measuring similar constructs** An essential element of the process of validation is correlating the test score(s) with scales from similar instruments, the so-called congruent or convergent validity. The guidelines on convergent 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 a time interval, lower values may be adequate. When evaluating convergent 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 convergent 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.	
	Not applicable	n/a
	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
10.2.1.5	**Do the correlations with other instruments show good discriminant validity with respect to constructs that the test is not supposed to measure?**	
	Not applicable	n/a
	No information given	0

	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
10.2.1.6	**If a Multi-Trait-Multi-Method design is used, do the results provide evidence for convergent and discriminant validity?** Note that if an MTMM design is used, research as mentioned in 11.2.1.3. and 11.2.1.4. may not be required anymore.	
	Not applicable	n/a
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
10.2.1.7	**(Quasi)experimental designs**	
	Not applicable	n/a
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
10.2.1.8	**Other** *(describe):* Click or tap here to enter text.	
	Not applicable	n/a
	No information given	0
	Inadequate	1

	Adequate	2
	Good	3
	Excellent	4
10.2.1.9	**Sample sizes** The guidelines below concern studies within the classical test theory framework. Note: If a sample has any characteristic that could justify its small size (e.g. clinical nature), please indicate it: Click or tap here to enter text.	
	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 adequately sized study	3
	Good range of adequate to large studies	4
10.2.1.10	**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. The selection of the marker tests is adequately justified and their psychometric properties are satisfactory.	4
10.2.1.11	**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	[ ]

10.2.1.12 **Overall adequacy based on the pattern and size of relationships and the adequacy of validation studies including sample size**

This overall rating is obtained by using judgment based on the ratings given for items in section 11.2.1. *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

10.2.2 **Relationships with a relevant criterion**

10.2.2.1 **Type of criterion study or studies**

Select all that apply.

Predictive	[ ]
Concurrent	[ ]
Post-dictive	[ ]

10.2.2.2 **Sample sizes**

Indicate whether any characteristics of the samples (e.g. clinical nature) could justify the small size of the sample(s): Click or tap here to enter text.

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 adequately sized study	3

	Good range of adequate to large studies	4
10.2.2.3	**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
10.2.2.4	**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.	
	Not applicable	n/a
	No information given	0
	Inadequate (AUC<65)	1
	Adequate (85 ≤ *AUC* ≤79)	2
	Good (80 ≤ *AUC* < 89)	3
	Excellent (*90* ≥ AUC≤100)	4

10.2.2.5	**Strength of the relation between the test and a qualitative criteria (diagnostic purpose)** In situations where sensitivity and specificity of a test is especially relevant (e.g. clinical and educational contexts) ROC-curves analyses are useful. In general, for the area under the curve (AUC), values between 90 and 100 are excellent, between 70 and 80 are acceptable, and below de 60. Swets (1988) presents an overview of values of ROC-curves in different areas. For certain types of medical diagnosis typical 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
10.2.2.6	**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	[ ]
10.2.2.7	**Evidence based on the relationship between the test and a criterion** This overall rating is obtained by using judgment based on the ratings given in section 11.2.2. *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 their 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

10.3. Evidence based on the internal structure (dimensionality) of the test

10.3	Evidence based on internal structure	
10.3.1	**Designs and/or techniques employed** Select all that apply.	
	No information is supplied	[ ]
	Exploratory Factor Analysis	[ ]
	Confirmatory Factor Analysis	[ ]
	Testing for invariance of structure and differential item functioning across groups	[ ]
10.3.2	**How 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: The results support the structure of the test both in terms of the number of factors extracted and their interpretation. In addition, sufficient and adequate information is provided to assess the quality of the decisions made in applying the technique - AFE and/or AFC, factoring method, rotation, software used, etc. - and to interpret the results.	4
10.3.3	**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: Detailed information on various studies on item bias related to gender, mother tongue, etc.). Use of appropriate methodology	4

10.4. Validity evidence based on other sources

10.4	Evidence based on other sources	
10.4.1	**Type of validity evidence** Select all that apply.	
	Based on cognitive or response processes	[ ]
	Based on consequences of testing	[ ]
	Other (please specify)	[ ]
10.4.2	**Adequacy of the studies that provide other sources of validity evidence** (in terms of sample size, methodological rigor, etc.)	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4
10.4.3	**How are the results of the studies providing other sources of validity evidence in terms of supporting the intended use of the test**	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent	4

10.5. Overall evidence validity

When judging overall validity, it is important to bear in mind the importance placed on different sources of evidence. 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.

10.5 Overall evidence validity

Validity – Overall adequacy
This overall rating is obtained by using judgment based on the ratings given for items 10.1,10.2, 10.3 and 10.4. *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).

Click or tap here to enter text.

11. Fairness, Diversity and Cultural Breadth

This section focuses on the fairness and appropriateness for use of the test with different demographic, cultural groups, differently abled and neurodiverse groups. An “inclusive by design” approach is preferred which considers the needs of different groups in the conception and development of the test. Whatever the design approach, information for the user regarding any differences that might be expected between groups, or adaptations that can be used with the test, is important to allow fairness in use. The reviewer should differentiate between tests used for diagnosis with particular groups (e.g. for identifying dyslexia) and the accessibility of a test designed for another purpose for a test taker from a particular group (e.g. a dyslexic student completing an interest inventory as part of careers guidance).

Items to be rated n/or 0 to 4		Rating					
11.1.1	**Rationale: Relevance of construct measured across groups and cultures** Excellent: Construct is defined in an inclusive manner with thought given to its applicability across all relevant groups and cultures.	n/a	0	1	2	3	4
11.1.2	**Documentation** Excellent: Full details of diversity and inclusivity issues considered during design and development and results of analyses. Procedural instructions are clear regarding the appropriate use of the test with people from different groups and provides advice on possible adaptations where needed and their impact on interpretation of results. c.f. 7.3.6; 7.3.7	n/a	0	1	2	3	4
11.2	**Development- Design** Excellent: Inclusive by design approach to developing tasks and items. Diversity in item writer and reviewer group. Built in adaptations where needed (e.g. computer delivery is compatible with voice reader technology, test taker can adapt font size and colour). c.f. 8.2.6	n/a	0	1	2	3	4
11.3	**Development – Piloting and Analysis** Excellent: Piloting with broad range of groups with particular emphasis on those that might be anticipated to have difficulty with the test/task (e.g. non-native speakers with a test with a large language component). Differential item functioning analysis for large range of groups as well as other item analysis informs item selection.	n/a	0	1	2	3	4
11.4	**Reliability** Excellent: Reliability considered for subgroups as well as majority group and test show similar levels of	n/a	0	1	2	3	4

	accuracy or precision for subgroups as for population as a whole.						
11.5	**Validity** Excellent: Construct findings robust across groups. Criterion validity studies using heterogenous groups. Actions taken to ensure equity in criterion measures. Differential validity studies performed where possible. Cf 11.1.4/11.1.5	n/a	0	1	2	3	4
11.6	**Interpretation**						
	11.6.1 Norm Referenced Interpretation Excellent: Norm groups reflect the diversity in population through appropriate sampling methods and over-sampling of small groups with particular diversity characteristics. Information provided about group differences and any differences found do not impact fair use of the test. c.f. 9.1.2/9.1.7	n/a	0	1	2	3	4
	11.6.2 Criterion-referenced Interpretation Excellent: The determination of the criterion score takes account of equality, diversity, and inclusion (EDI) considerations. Any reviewers used in standard setting have training in EDI considerations and where possible reflect the diversity of the test taker group.	n/a	0	1	2	3	4
11.7	**Reports** Excellent: Reports are free from bias and written using inclusive language. Where group differences are anticipated, appropriate guidance is provided for users of reports. Interpretations and recommendations are appropriate for all relevant groups and avoid stereotypical thinking. c.f. 12.4	n/a	0	1	2	3	4
11.8	**Overall Evaluation** The overall evaluation is obtained using judgement based on the ratings given for items 11.1-11.7. Do not simply average numbers to obtain an overall rating.						
	No information given	0					
	Inadequate	1					
	Adequate	2					
	Good	3					
	Excellent	4					

12. Quality of digitally generated reports

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.

'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	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?*	
	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 bias** - *Is the content of the report and the language used biased against certain groups?* - *Does the report make clear any areas of possible bias in the results of the instrument?*	
	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	**Level of Detail** Reports can be very short and lack detail, missing out relevant information that can be derived from test scores. They can also be overlong which can be an indication of overinterpretation of scores with detailed descriptions that cannot be supported by test scores alone. As a rule of thumb, reports that on average take more than one page per scale (excluding title pages) may be overlong and overinterpreted.	
	No information given	0
	Inadequate	1
	Adequate	2
	Good	3
	Excellent: Level of detail is appropriate to the test and the purpose of the report. The reader can easily see the most important information which is presented with at a suitable level of detail. It is not obscured by too much detail.	4
12.7	**Overall adequacy of computer-generated report** This overall rating is provided for each report and 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. Also where there are multiple reports a comment on the consistency of quality of reports.

Click or tap here to enter text.

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.

Include any positive or negative points raised in connection with adapted and translated tests. 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. Include comments on any research that is known to be under way, and the supplier's plans for future developments and refinements etc.

Click or tap here to enter text.

Conclusions

Click or tap here to enter text.

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	1) Requires further development before being used.	[ ]
	2) Only suitable for use by an expert user under care- fully controlled conditions or in very limited areas of application	[ ]

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.	3) Suitable for supervised use in the area(s) of ap-plication defined by the distributor by any user with general competence in test use and test admini-stration	[ ]
	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	[ ]
	5) Suitable for unsupervised self-assessment in the area(s) of application defined by the distributor	[ ]
	6 Other: Click or tap here to enter text.	[ ]

Part 3. Bibliography

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Bartram, D. (2005). The great eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology*, 90(6), 1185–1203.

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.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, 111(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bugbee, A. C. (2014). The Equivalence of Paper-and-Pencil and Computer-Based Testing. *Journal of Research on Computing in Education*, 28(3), 282–299. https://doi.org/10.1080/08886504.1996.10782166

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

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.

Egberink, I.J.L. & Leng, W.E. de. (2009-2024). *COTAN Documentatie* (www.cotandocumentatie.nl). Amsterdam: Boom Uitgevers Amsterdam.

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

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., et al. (2012). Testing practices in the 21st Century: Developments and European psychologists’ opinions. *European Psychologist*, 17(4). p.300-319.

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

Gisev, N., Bell, J.S., & Chen T.F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, Volume 9, Issue 3, 330-338. https://doi.org/10.1016/j.sapharm.2012.04.004

Hagemeister, C., Kersting, M., & Stemmler, G. (2012). Test Reviewing in Germany. *International Journal of Testing*, 12(2), 185–194. https://doi.org/10.1080/15305058.2012.657922

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

Iliescu, Dragos (2017). *Adapting tests in linguistic and cultural contexts*. Cambridge University Press, Cambridge

Innocenti, F., Tan, F. E. S., Candel, M. J. J. M., & Van Breukelen, G. J. P. (2023). Sample size calculation and optimal design for regression-based norming of tests and questionnaires. *Psychological Methods*, 28(1), 89–106. https://doi.org/10.1037/met0000394

International Test Commission (2017). *The ITC Guidelines for Translating and Adapting Tests* (Second edition). [www.InTestCom.org]

King, W. C., & Miles, E. W. (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology,* 80, 643–651.

Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J *Clin Epidemiol. 2011* Jan;64(1):96-106. doi: 10.1016/j.jclinepi.2010.03.002

Landis J.R., Koch G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics*; 33: 159–174.

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.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114(3), 449–458. https://doi.org/10.1037/0033-2909.114.3.449

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.

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

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

Oosterhuis, H. E. M., van der Ark, L. A., & Sijtsma, K. (2016). Sample Size Requirements for Traditional and Regression-Based Norms. Assessment, 23(2), 191-202. https://doi.org/10.1177/1073191115580638

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

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

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.

Timmerman, M. E., Voncken, L., & Albers, C. J. (2021). A tutorial on regression-based norming of psychological tests with GAMLSS. *Psychological Methods*, 26(3), 357–373. https://doi.org/10.1037/met0000348

Trahan, Lisa H.; Stuebing, Karla K.; Fletcher, Jack M.; Hiscock, Merrill (2014). "The Flynn effect: A meta-analysis". *Psychological Bulletin*. 140 (5): 1332–1360. doi:10.1037/a0037173. ISSN 1939-1455. PMC 4152423. PMID 24979188

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.

Contact

European Federation of Psychologists' Associations AISBL- EFPA

Rue Marché aux Herbes 105 / 39 – 1000 Bruxelles - Belgique
Grasmarkt 105 / 39 - 1000 Brussels – Belgium

headoffice@efpa.eu

Follow us

https://www.efpa.eu

https://www.facebook.com/europeanpsychology

https://be.linkedin.com/company/europeanpsychology