Nederlands Instituut van Psychologen NIP

COTAN review system for evaluating test quality

Arne Evers, Wouter Lucassen, Rob Meijer and Klaas Sijtsma

Contents

Authors
Arne Evers, University of Amsterdam, Work and Organisational Psychology
Wouter Lucassen, COTAN member
Rob Meijer, University of Groningen, Faculty of Behavioural and Social Sciences, Department Psychometrics and Statistics
Klaas Sijtsma, Tilburg University, TS Social and Behavioural Sciences, Department of Methodology and Statistics

Completely revised version, May 2009, reprinted 2010, with the addition of revised tables for establishing final rating of a criterion dated 1 July 2015

Preface

Here we present the revised version of the COTAN Review System for Evaluating Test Quality. This version is based on previous versions of the system, as published in the *Documentatie van Tests en Testresearch* of 1982 (Documentation of Tests and Test Research: Visser, van Vliet-Mulder, Evers & ter Laak, 1992) and 2000 (Evers, van Vliet-Mulder & Groot), and on the NIP website in 2004 (www.psynip.nl). This new version was prepared by a work group of the Dutch Committee on Tests and Testing (COTAN), comprised of the four authors of this preface. All COTAN members contributed to the discussion on the content of this new review system, and this text was approved at the COTAN meeting of 19 March 2009. At the time of this revision, the COTAN was made up of the following members: K. Sijtsma (chair), J.B. Blok (secretary), A. Evers (senior editor test reviews), R.M. Frima (staff member), R.H. van den Berg, M.Ph. Born, H.W. van Boxtel, M.E. Dinger, B.T. Hemker, P.P.M. Hurks, W.W. Kersten, W.I. Lucassen, R.R. Meijer, E.F.M. Pouw, W.C.M. Resing, J.D.L.M. Schutijser and T. van Strien.

This revision involves major changes to the review system in response to the developments of recent years that have taken place in the areas of test theory and test construction. Examples of these developments include *computer-based tests, item-response theory* and *continuous norming*. Although the previous version of the review system could be applied with some flexibility to tests that employed these new developments, the text was primarily geared towards tests that were developed and administered in the classical manner. In this revision, such new techniques and approaches are more explicitly addressed. In this way, the COTAN hopes to bring the review system to a new level of utility and applicability. The complexity of the new developments, as well as the necessity of properly reporting the employed procedures and results to test users and to the COTAN, makes high demands on the knowledge and skills of test authors. When applying these more advanced techniques, test authors will find it more often necessary to call on the expertise of specialists.
The new system rates tests on the seven criteria familiar from the previous version, which are stated in the introduction to this text. It goes without saying that advancing insights into test theory, test construction and test application have led to modifications in the questions used for evaluation and the way in which these lead to a rating for each criterion. One consequence of this is that the rating results for a particular instrument may show differences depending on whether the old or the new system is used.

Has the review system been made stricter? In general, this is not the case. The previous system evaluated certain features quite harshly which can now be judged with more nuance, but on other points the requirements have been further tightened. However, the most important change is that the review procedure has become much more specific, not necessarily stricter or more lenient.

This publication gives a concise introduction to the quality requirements for the various criteria. These observations make no claim to comprehensiveness, nor do they constitute a recipe book for test construction. It is reasonable to expect test authors to be familiar with the fundamentals of psychometrics and current quality standards in test construction, so that they can take full responsibility for the quality of their instrument.

This revision incorporates passages by other authors, with their consent: Keuning (2004) on computer-based tests; Wools, Sanders and Roelofs (2007) on absolute norming; and Bechger, Hemker and Maris (2009) on continuous norming. The COTAN owes these authors a debt of gratitude for their expert contributions. The second printing has incorporated the errata in the table for establishing final ratings for criterion 4 (Norms). A few small editorial corrections have also been made.

Arne Evers, Wouter Lucassen, Rob Meijer and Klaas Sijtsma
April 2010

Comments on the translation

March 2019

This translation was commissioned by the Dutch Committee on Tests and Testing (COTAN). Currently, the Committee is working on a revised version of the COTAN Review System for Evaluating Test Quality. This revised version will also become available in English.

There have been some developments since the publication of the Dutch version of the COTAN Review System for Evaluating Test Quality in 2010, that we would like to bring to your attention:

- A new version of the *General Standards for Test Use* was published in May 2018, called *the Guidelines for the Use of Tests 2017*. These guidelines are available on the website of the Dutch Association of Psychologists (www.psynip.nl).
- In 2016, a new version of the online COTAN Documentation was released (Dutch only). In this database all COTAN reviews are published.
- The COTAN published the following addenda (Dutch only) to the COTAN Review System for Evaluating Test Quality:
  - Addendum on fairness (2015)
  - Addendum on unproctored data collection (2018)
  - Addendum on the updating of tests (2018)
- Since 2010, the COTAN published the following additional documents (Dutch only):
  - Guidelines for a response to an initial test review of the COTAN (undated)
  - General Terms and Conditions (2014)
  - Guidelines on impartiality (2018)
- Up-to-date information on the COTAN review procedure is published on the NIP website in Dutch and summarized in English. There you can also find the *Guidelines for the Use of Tests 2017*, the COTAN addenda to the COTAN Review System for Evaluating Test Quality (2010) and the additional documents published by the COTAN.

Introduction

Contents of the review system

Although the COTAN has been describing test contents ever since its founding in 1959, it was in their *Documentation of Tests and Test Research* (1969) that they first published their ratings of tests. The system used led to a general evaluation of tests. The rating could vary from A (excellent test) to F (poor test or test under development). Keeping in mind the need for creating a more multi-faceted rating, the *Documentation* of 1982 used an entirely new system. Each test was rated with five criteria: 1) theoretical basis of the test construction, 2) quality of the test materials and test manual, 3) norms, 4) reliability and 5) validity. When assembling the 2000 edition of *Documentation of Tests and Test Research* the system was revised, and the questions, explanations and instructions for weighting were modified. This involved for the most part relatively small adjustments and improvements. Two substantial changes, however, were the splitting of criterion 2 into separate ratings for quality of the test materials and for quality of the test manual, and the splitting of criterion 5 into construct validity and criterion validity. At the time, it took little effort to convert the ratings on five criteria into more specific ratings of the seven new criteria.
In the current revision, however, the alterations are more substantial, although the number of criteria has remained at seven. Similarly, the principle of various questions per criterion, including one or more key questions, has remained unchanged. However, new insights in the areas of test theory and test construction have meant that some questions have been added, while other questions have been divided into sub-questions, and still others have been removed. Additionally, instructions have been specified and scoring rules have been changed. If one uses both the old and the new system to evaluate a test, this can of course lead to different reviews. As of April 2009, tests have been evaluated using the new system. No re-evaluations will take place of the approximately 600 tests that have been evaluated with the old system. The only exception to this concerns norm data (see question 4.2) which have become outdated.

Here follows a general description of the revised review system, including the most significant changes. The seven criteria are:

1. ***Theoretical basis of the test construction***
   This criterion is rated using questions which determine whether the content of the test reflects its intended purpose, its theoretical background and its operationalisation, in that order. The rating of this criterion has consequences for the rating of other criteria because the choice of constructs for measurement dictates what type of research must be performed for norming, reliability and validity. The greatest departure from the previous version of the review system is that the first question is split into three sub-questions which explicitly ask whether the intended construct or constructs, target group or target groups and function of the test are described.
2. ***Quality of the test materials***
   This criterion is rated with eight questions. This criterion covers such matters as whether test items, scoring and instructions are standardised, and whether sufficient directions are provided on how to take the test. There is also a question about content that is possibly offensive to specific subgroups in the population. New questions in this criterion include a question about the quality of the items, and separate question series for administration with paper-and-pencil or by computer.
3. ***Quality of the test manual***
   This criterion is rated either with seven paper-and-pencil questions, or ten computer questions. This criterion enquires about the information supplied to support test users as they administer and interpret the test. The main change is that three additional questions are added for computer tests.
4. ***Norms***
   This criterion is rated with seven questions (norm-referenced interpretation) or five questions (content-referenced or criterion-referenced interpretation). One new aspect for all types of norms is that the review imposes a time limit for outdated norms. For norm-referenced interpretation the quality of the norms, and of the additional information is evaluated. For norms that are calculated with continuous norming, target figures are now given for the desired size of norm groups. Also new are the questions concerning content-referenced and criterion-referenced interpretation.
5. ***Reliability***
   This criterion is evaluated with three questions. The size of the reliability coefficients is evaluated first, followed by the quality of the research carried out on the reliability. As a result of new developments, information on six possible reliability indices is requested, in contrast to four in the previous version of the review system.
6. ***Construct validity***
   This criterion is evaluated with three questions. The outcomes are evaluated first, followed by the quality of the research carried out on the construct validity. New are more explicit statements of what sort of research data can serve to support the construct validity, and what types of data are required for a particular rating.
7. ***Criterion validity***
   This criterion is also evaluated with three questions. As with construct validity, the first question is about the outcome sizes, after which these are evaluated in the light of the quality of the research procedure. No significant changes have been made to this criterion.

The rating for each of these criteria can be 'insufficient', 'sufficient' or 'good'. The scoring scale for the questions is 1, 2, and 3; these correspond with the designations 'insufficient', 'sufficient' or 'good'. In a few questions, a score of '1' is to be interpreted as 'no', 2 as 'not applicable' and 3 as 'yes'. A negative rating of a key question or its sub-question immediately leads to a rating of 'insufficient' for the criterion in question. The content of the questions, the explanatory notes and the weighting rules are discussed in the next seven chapters. According to the weighting rules for some criteria, sum scores of questions must be calculated.
The purpose of the explanatory notes is to assist the review process and to clarify the statistical or psychometric motivations, when necessary. The explanatory notes obviously make no claim to be a statistical or psychometric textbook. In case of confusion, it is best to consult the references or other literature in the areas of test construction and psychometrics.

The meaning of the ratings

Generally, we can state that there are two ways to obtain an insufficient rating for a criterion: either the required information is lacking, or the quality of the information that is available receives a negative rating. For example, an 'insufficient' score for test reliability can mean that either the reliability was not investigated or that it was investigated, and that the research has revealed that the test's reliability is insufficient. Absence of research data is therefore evaluated in the same way as research data that are available but lead to a negative result, since the COTAN assumes that it is the author's duty to provide research data. In this way, we follow the scientific custom that the burden of proof for a statement lies with the researcher. For the example described above, this means that a test for which there are no data is seen as insufficiently reliable until proven otherwise. For test users, it can be useful to differentiate between these situations, for example if one wants to give a new and promising instrument the benefit of the doubt. In order to enable this distinction, but also to supply an extra information source to test authors and users, 'insufficient' scores for tests rated since 1992 have included a brief explanation of the reason for such ratings. Once again, this underscores the responsibility of the test author to provide sufficient information at the appropriate time. On the other hand, the appraisal users give of an insufficiently substantiated instrument must be commensurate with that instrument's age.

A second refinement of the 'insufficient' rating is that one or two 'insufficients' do not automatically render an instrument unusable. For example, an 'insufficient' for norms might be given because the norm group is not representative enough. However, the test may still be usable if the user is able to assemble suitable norms on his own. Similar considerations apply to reliability and validity. One or more scales or subtests in a questionnaire or test may be insufficiently reliable, but this does not necessarily indicate that the other scales or subtests, or the total score, are unusable. Tests used for important decisions at an individual level are subject to especially high standards of reliability (see the notes to criterion 5). The reliability of such a test is rated as 'insufficient' if the reliability coefficient is less than .80. Admittedly, such a test can still yield useful information when it is combined with other instruments, but since the present review system can be used only for individual tests, this type of use does not apply here. Within this review system, it is also possible that the same test may receive an 'insufficient' for reliability, while the construct validity or criterion validity may be found 'sufficient' or even 'good'. For example, a validity coefficient of .40 is rated as 'good' in selection situations. In some cases, even a test with a low predictive value can yield useful information, depending on factors such as base rate, selection ratio and cost-benefit ratio.

A third refinement is the cut-off values stated in the review system, to which tests must conform in order to guarantee the maximum objectivity of the rating. For example, the criteria for norms and reliability state specific sample sizes and reliability coefficient sizes, respectively, which must be achieved for a rating of 'sufficient' or 'good' and serve as an anchoring point for the reviewer. It is true that no decisive scientific arguments can be offered for these levels: they are based on generally accepted recommendations from prominent scholars (see the relevant chapters for references). As a result, for every case of values that lie close to these limits, it is very difficult to argue why one particular value narrowly merits a 'sufficient' or 'good', while another value narrowly falls short. All the same, this approach better guarantees that all tests are essentially reviewed in the same way.

The above remarks are intended to make it clear that the test user is expected to be able to deal correctly with the absolute terms used to describe the ratings. For expert test users, the rating of 'insufficient' for any criterion whatsoever functions primarily as a warning sign. In such a case, the test user must make use of article 3.2.e of *Algemene Standaard Testgebruik (General Standards for Test Use)* (Dutch Association of Psychologists, 2004), to explicitly state why he is employing the instrument in question. For less experienced users, especially when a test has received multiple insufficient scores, the message is, 'User, avoid this test!'

Review procedure

It is sometimes thought that the COTAN reviews only those tests that have been submitted for that purpose. This is not the case, because the COTAN functions in a proactive manner. In principle, the Committee evaluates all tests under consideration for inclusion in the *Documentation for Tests and Test Research*. (These inclusion criteria are given on the NIP/COTAN website under the heading of Criteria, see www.psynip.nl or www.cotan.nl.) The first step, certainly when the COTAN itself takes the initiative in reviewing a test, is to assemble all materials and publications for and about a test: test booklets, keys, manuals, software, articles and dissertations, and so forth. These materials and publications are normally supplied by the authors or publishers, either spontaneously or upon request. One problem group is authors who refuse for one reason or another to release the test materials for reviewing (see the following section). These tests can neither be documented or reviewed.

The assembled material is sent to two reviewers who work independently of one another. In accordance with the policies of recognised international psychology journals, the reviewers of a specific test remain anonymous. All COTAN members, as well as a group of external experts approached for this purpose, serve as reviewers. A reviewer is assigned a test on the basis of his expertise in a particular area. Furthermore, the intention is that at least one of the two reviewers is a COTAN member. Care is also taken that reviewers will never evaluate a test constructed either by themselves or a direct colleague, nor by a competing organisation. In case of disagreement in the ratings, the reviewers are requested to mutually arrive at a consensus. In exceptional cases, a third reviewer is brought in. Fundamental issues concerning ratings are discussed at the bimonthly COTAN meetings.

After both reviewers submit substantive explanation of their ratings, the senior editor integrates this into a report which is sent to the test author as feedback, along with the final ratings. The test author then has the opportunity to react to the ratings. Reactions and comments from the author are dealt with by the reviewers. If required, subsections of tests can be evaluated for a second time, possibly by an independently operating third reviewer. The review is then published. To this end, the online version of the *Documentation of Tests and Test Research* is updated on a monthly basis. With the implementation of this revised version of the review system, it was also decided to publish the integrated commentary supplied by the reviewers, since this can help explain the reasons for a particular rating. An attempt will be made to publish this commentary as well for tests that were recently evaluated using the 'old' version of the system.

The above passage has described the review procedure for a new test. When new information on a previously evaluated test appears, such as a revised manual, research report or complete revision of the test, a re-evaluation of the test may take place. The procedure described above is then repeated in its entirety. One limit is in force: at least one year must have passed since the previous evaluation of the test.

Confidentiality

The review procedure incorporates two forms of confidentiality. As stated above, the first of these is the anonymity of the reviewers. The test author remains unaware of who has evaluated the test. Correspondence proceeds through the COTAN documentation centre or through the COTAN's senior editor for test reviews. This prevents discussions of a review from playing out on a personal level. This procedure is in agreement with that used by major psychological journals for the review of manuscripts. However, once a year, a list of reviewers is published on the COTAN website with the number of tests they have evaluated.

The second form of confidentiality concerns the collected or supplied test materials. Third parties have no access to information contained in the COTAN documentation centre and the materials are sent to reviewers under the condition of absolute secrecy. The COTAN is thoroughly aware not only that tests are subject to authors' rights, but also that a test is generally the end product of a costly developmental process and constitutes the working capital of psychologists and consultancy firms, which they use to distinguish themselves from others in their field. For this reason, the review procedure is carried out meticulously, and the test author can rest assured the test materials will not be accessed by parties other than those involved with the evaluation of the test in question. This aspect is emphasised here because the primary reason that test authors give for not providing test materials is the fear that this places the test in the public domain. There is no basis for this concern. The only information about the test that is made public is the standard description in the *Documentation*, the ratings of the seven criteria according to the system, and the commentary of the reviewers.

To protect a test against unauthorised copying, some authors and publishers have resorted to limiting the amount of information contained in the manual. This concerns mostly information on the content of subscales and information on norming. As a result, the test user does not know which items belong to which scale, or perhaps cannot even consult the table of norms. In such cases, the scoring is performed with online software or at a distance, by the publisher's own scoring service. The COTAN is most emphatically not an advocate of this practice, as it seriously impedes the test user's options for interpretation. It is important to the review that the reviewers at least have access to all the information that would enable a complete review of the test. In such cases, the author or publisher is requested to provide such information for the review. It goes without saying

that this is treated with the same degree of confidentiality as described above. The lack of such information can result in a rating of 'insufficient' for one or more criteria.

Translations or adaptations of foreign tests

Many Dutch tests are translations or adaptations of foreign instruments. When evaluating these instruments, one may ask to what degree research performed with the foreign version of the test is applicable to Dutch settings or can play a part in the evaluation of the Dutch version. The answer to this question depends in part on how literally the test has been translated. In some cases, the aim is to translate the foreign instrument as literally as possible into Dutch, both in linguistic terms and with agreement in meaning of items and concepts. The *International Test Commission* states a number of guidelines for this purpose in *Test Adaptation Guidelines* (ITC, 2000; see also Van de Vijver & Hambleton, 1996, and Hambleton, Merenda, & Spielberger, 2005).

In the first place, the guidelines address the way in which the translation (including back translation) is made. An essential point is that the translation need not be absolutely literal, and that it is far more important that the texts should sound natural. This will better ensure that the meaning of what is being asked is the same in both languages. Subsequently, equivalence research must provide 'proof' that the translated version is measuring the same construct as the original version. One example of this is the comparison of the factor structure of the original and Dutch versions. Once equivalence is established, the validity data and any test-retest data of the original version can be included in the evaluation of the Dutch version. Data in the sense of internal consistency can be calculated on the basis of the new Dutch norm data that must still be collected. The author of the Dutch version must provide a summary of any relevant foreign research in the manual; a mere reference is not sufficient.

In some cases, the Dutch version is clearly an adaptation of the original foreign test. This may be because some items are unsatisfactory and have had to be replaced, because the test has been expanded, because the answer scale has been altered, or because the notes and instructions have been changed. In such cases, the research data collected abroad may not take the place of Dutch research. If the intended construct to be measured is not altered and the same constructs in general are being measured, one may assume that the Dutch version of the instrument has the same theoretical basis as the original test. However, this must also be stated in the Dutch manual.
In none of the above cases may norm data from the foreign instrument be generalised to the situation in the Netherlands. These will have to be newly collected here.

Dutch-language tests from Belgium

The *Documentation* includes a number of Dutch-language tests originating from Belgium. This origin is always stated as part of the review. These tests are evaluated and included in the *Documentation* because in principle they can be administered in the Netherlands without translation or modification. Of course, the evaluation takes place for use of the instrument in the Netherlands. The following rules have been applied in this process:

- ***Theoretical basis of the test construction***
  The evaluation of this criterion is affected only when the test has been constructed for a specifically Belgian situation, for example a test for interest in a school system that does not exist in the Netherlands. The rating in this case is 'insufficient'.
- ***Quality of the test materials***
  The Flemish language variant uses words, idioms and sentence constructions different to those in Dutch. Since the test is being evaluated for use in the Netherlands, it is a condition that all tasks and instructions must be stated in standard Dutch. Questions 2.6 and 2.14 can be used to help answer this question.
- ***Norms***
  When norms are collected with the aid of Dutch groups, these are evaluated in the normal fashion. Norms that are based on Belgian groups receive a rating of 'insufficient', unless it is deemed credible that the score distribution for these groups is similar to that of comparable Dutch groups.

No special requirements apply to reliability and validity. It is assumed that these data are generalizable to the Dutch situation, unless there are explicit reasons to assume otherwise.

In fact, Dutch-language tests from Belgium are treated, at least partially, in the same way as other tests of 'foreign' origin. These reviews can of course not be used as an indicator of the quality of the test in question for use in Belgium.

Summary

This section summarises the principal subjects that have been treated in this introduction.

- The test review focusses on the following seven criteria: 'Theoretical basis of the test construction', 'Quality of the test materials', 'Quality of the test manual', 'Norms', 'Reliability', 'Construct validity' and 'Criterion validity'.
- The rating for each of these criteria can be 'insufficient', 'sufficient' or 'good'.
- The ratings made with the revised version of this review system will not entirely match the ratings made with the 'old' system.
- The rating 'insufficient' can mean that certain information is unavailable.
- The COTAN guarantees complete confidentiality of the test materials supplied by the author or publisher.
- All relevant information for the review must be supplied to the COTAN reviewer.
- For adaptations or translations of foreign tests, a description of the 'theoretical basis of the test construction' must be included in the manual.
- Generalisation of research findings from foreign tests to the Dutch version is only possible if the two versions have been shown to be equivalent.
- Flemish-Belgian tests are explicitly evaluated for use in the Dutch situation on the criteria 'Theoretical basis of the test construction', 'Quality of the test materials' and 'Norms'. These reviews are not applicable to use in Belgium.

1 Theoretical basis of the test construction

Test construction demands thorough preparation. After all, test users need to make sound judgements in many sensitive situations. These might concern intra-individual differences within such contexts as educational follow-up systems where time differences can play a role, or the use of ipsative interest inventories in vocational guidance, inter-individual differences as in personnel selection, and differences between groups or situations as in organisational research. The information provided by the test author should enable the prospective test user to judge whether the test is suitable for the required purpose. There must thus be a clear description of the construct the test intends to measure as well as a detailed justification of the choice of test content and the manner in which the constructs are to be measured.
This criterion is concerned only with establishing that the basic principles have been explicitly stated. The quality of the research design and execution of research is not relevant here; these are treated in criteria 3 through 7.

Notes for key question 1.1: 'Is the purpose of the test specified?'

a. The construct or constructs being measured must be specified. It must thus be clear what constructs the test is meant to measure. In this case, 'construct' has a broad meaning and can refer to a domain of specific behaviours, preferences or styles. It is vital to properly define which behaviours belong to the domain being measured. The construct being measured might therefore be intelligence, reading skills, motivation to perform, job interests or ADHD.
b. The target group or groups for the test must be specified, in such terms as age, profession, educational level, relevant prior knowledge or normal versus clinical contexts. The more extensive the claim of broad applicability, the greater the obligation to supply empirical materials such as norms or validation data.
c. Test construction begins with a reflection on the purpose of the test. Is the aim to predict criterion behaviour? Is a test intended to measure educational progress or the effects of training? Is it being used to assess the levels of students for allocation purposes? Is it intended for diagnosis leading to a treatment plan?

Notes for question 1.2: 'Is the source of the test construction described, and/or are the constructs that the test purports to measure clearly defined?'

Does the test reflect an existing theory, or has the author developed a new one? Is this theory adequately described? If the test is a translation or an adaptation of a foreign instrument, background information on that instrument must be provided: a list of references alone is not enough. Even when a test is intended for the measurement of well-known constructs such as intelligence, a definition of the construct must be given to clarify exactly which facets of behaviour belong to the domain. When a historically based test method (as opposed to one with a theoretical foundation) is used to measure a construct in a traditional way, arguments must be given to show why it is useful to measure these constructs. Similarities and differences with comparable tests must also be described. What is the added value of the new instrument over existing instruments? When the test is a variant on pre-existing instruments or an adaptation of the paper-and-pencil version for computer use, are the differences between these two versions noted and specified?

Questions for criterion 1
Theoretical basis of the test construction

		ins.	suf.	good
Key question 1.1	Is the purpose of the test specified? a. Is it clearly specified what construct or constructs the test intends to measure? b. Are the target group or groups for the test specified? c. Is the function of the test specified? If the rating for one or more of the sub-questions a, b or c is negative (1), skip the other questions for this criterion and continue with criterion 2.	 1 1 1	 2 2 2	 3 3 3
1.2	Is the source of the test construction described, and/or are the constructs that the test purports to measure clearly defined?	1	2	3
1.3	Is the relevance of the test content for the construct or constructs to be measured justified?	1	2	3

Notes for question 1.3: 'Is the relevance of the test content for the construct or constructs to be measured justified?'

This question applies to the step from intended measurement to operationalisation. Is a definition of the content domain supplied which makes clear whether or not a particular item belongs to the test? Are the constructs to be measured analysed in such a way that it becomes clear what facets can be distinguished within the construct(s)? If needed, are theoretical or content-related considerations used to assign different weights to these facets, and is this taken into account when choosing the items? If items are deleted or modified in the course of constructing or adapting a test, does the author indicate the consequences of these changes for the measurement of the original construct? In other words, is the content domain still completely covered, or has it narrowed or shifted?

In adaptive tests (those in which the choice of items offered to the test taker is determined on the answer pattern they have shown on items answered thus far), is it specified how the test content is to be guaranteed? In adaptive tests, each candidate is presented with a different set of items, which means that certain content may be under-represented in the test. For this reason, it is often necessary to perform a content check, using a method such as that of Kingsbury & Zara (1991), so that every test is in agreement with the specification table.

Rules for determining final rating for criterion 1
Theoretical basis of the test construction

The sum of the ratings for questions 1.1.a to 1.1.c is 8 or 9.	Both of the other questions are rated at least '2'.	good
	One of the two other questions is rated '2' or '3' and the other is rated '1'.	sufficient
	Both of the other questions are rated '1'.	insufficient
The sum of the ratings for questions 1.1.a to 1.1.c is 6 or 7 and none of these questions are rated '1'.	Both of the other questions are rated at least '2'.	sufficient
	One or both of the other questions are rated '1'.	insufficient
One or more of questions 1.1.a to 1.1.c are rated '1'.		insufficient

2 Quality of the test materials

When rating this criterion, a distinction is made between tests using paper and pencil and those administered with the use of a computer (computer-based tests: CBT). In CBT mode, there is no distinction between tests taken in local test environments and those taken via the internet, because the requirements are essentially the same.

In this criterion, three key questions are asked for both testing modes. In general, meaningful interpretation of a test score demands that the test must be administered and scored in such a way that no unintended factors can exert influence on the calculation of the score. The administration and the instruction must be standardised in such a way that the influence of varying instructions or different administrators (for paper tests) or differing test situations or decision rules (for an adaptive test) is eliminated, or in any case kept as limited as possible. The score must also be as objective as possible.

A score of 2 for the third key question on content that is racist or offensive to certain minority groups can lead to a separate remark for the rating such as 'very restricted usability for ethnic minority groups' (see Hofstee et al., 1990).

When rating an instrument intended for paper-and-pencil use, begin with question 2.1; when the instrument is intended for computer use, begin with question 2.9. If an instrument exists in both paper-and-pencil and CBT form, the quality of the test materials for both forms must be rated. If the ratings disagree, this can be indicated in a footnote to the rating. The assumption here is that the items and, as far as possible, the instructions for both versions are identical. If this is not the case, we are in fact dealing with two different tests, and a complete review of both versions must be separately prepared. In this case, the psychometric data will definitely not be generalizable between the versions.

Questions for criterion 2
Quality of the test materials
Paper-and-pencil version

		ins.	suf.	good
Key question 2.1	Are the test items standardised? If the rating of this question is negative (1), skip the other questions for this criterion and continue with criterion 3.	1		3
Key question 2.2	a. Is there an objective scoring system? and/or: b. In case the test has to be scored by raters or observers, is there a clear and complete system for rating or observation? If the rating of this question is negative (1), skip the other questions for this criterion and continue with criterion 3.	1 1	2 2	3 3
Key question 2.3	Are the test items free from racist, ethnically directed and sexist content, or any other content offensive to specific subgroups in the population? If the rating of this question is negative (1), skip the other questions for this criterion and continue with criterion 3.	1	2	3
2.4	Are the items, test booklets, answer scales, and answer sheets devised in such a way that filling-in errors can be avoided?	1	2	3
2.5	Are the instructions for the test taker complete and clear?	1	2	3
2.6	Are the items correctly formulated?	1	2	3
2.7	What is the quality of the test materials?	1	2	3
2.8	Is the scoring system devised and explained in such a way that scoring errors can be avoided?	1	2	3

Questions for criterion 2
Quality of the test materials
Administration by computer

		ins.	suf.	good
Key question 2.9	Is the test standardised, or are decision rules for adaptive tests specified? If the rating of this question is negative (1), skip the other questions for this criterion and continue with criterion 3.	1		3
Key question 2.10	Is an automated or objective scoring system in force? If the rating of this question is negative (1), skip the other questions for this criterion and continue with criterion 3.	1	2	3
Key question 2.11	Are the test items free from racist, ethnically directed and sexist content, or any other content offensive to specific subgroups in the population? If the rating of this question is negative (1), skip the other questions for this criterion and continue with criterion 3.	1	2	3
2.12	Has the software been designed in a way that prevents errors caused by incorrect use?	1	2	3
2.13	Are the instructions for the test taker complete and clear?	1	2	3
2.14	Are the items correctly formulated?	1	2	3
2.15	What is the quality of the design for the user interface?	1	2	3
2.16	Is the test sufficiently secured?	1	2	3

PAPER-AND-PENCIL VERSION

Notes for key question 2.1: 'Are the test items standardised?'
Test items are standardised when they are the same for every respondent with respect to content, form and order. This is an important condition for interpreting and comparing scores. An exception with respect to the requirement of a uniform order of test items is made for adaptive tests (see question 2.9). Although adaptive tests almost always come in CBT form, some paper-and-pencil tests also have adaptive features, such as rules for starting and breaking off the test.

Notes for key question 2.2.a: 'Is there an objective scoring system?'
A scoring system is called objective when the score values assigned to all possible answers by the participants are sufficiently agreed upon such that any qualified person who scores the items will, apart from administrative errors made either in hand or automated scoring, give exactly the same score. This is particularly applicable to paper-and-pencil ability tests and questionnaires with multiple-choice items.

Notes for key question 2.2.b: 'In the event that the test has to be scored by raters or observers, is there a clear and complete system for rating or observation?'
For observation scales, projective tests, subtests of individual intelligence tests with open questions, and essay questions, scoring cannot be strictly objective. However, procedures must be described that will guarantee maximum objectivity. This means that guidelines must be stated for rating and scoring, including model answers, model behaviours, scale anchors, instructions for weighting and so forth. These must clarify exactly what an answer must contain or what behaviour must be displayed for a particular score to be awarded. If applicable, the nature and content of the training that the raters have received must also be described.

Notes for key question 2.3: 'Are the test items free from racist, ethnically directed and sexist content, or any other content offensive to specific subgroups in the population?'
In 1990 a task force jointly established by the Dutch Association of Psychologists and a national anti-discrimination bureau screened 20 of the most frequently used Dutch tests for possible racist content (Hofstee et al., 1990). Although none of these tests were found to contain racist content, ethnically directed comments and unnecessary idiomatic expressions were found, especially in the verbal tests. Tests containing this material were assigned the label 'very restricted usability for minorities'. This strategy has been adopted for the present review system. The principle of restricted usability can also apply to other groups, for example an interests inventory that shows exclusively pictures of men. In such cases, this question must be rated with '2'. Explicitly racist or sexist content always renders a test unusable for all groups (score of '1'), except if this content matches one of the constructs being measured, such as the F scale or the androgyny scale.

This question is not intended to determine whether bias research has been performed (this falls under criterion 6), nor does it assess the items for bias. It simply has to do with the basic usability of a test for a particular group.

Notes for question 2.4: 'Are the items, test booklets, answer scales, and answer sheets devised in such a way that filling-in errors can be avoided?'
Points which must be addressed when rating this question are:
- The questions or items must be comprehensible (thus not too difficult) for the intended test group.
- If a separate answer sheet is being used, it must be designed in such a way that mistakes like skipping an item can be avoided and that any mistakes made by the test taker can be quickly detected.

Notes for question 2.5: 'Are the instructions for the test taker complete and clear?'
One must differentiate between the instructions for the test taker and the test administrator. The quality of the instructions for the test taker are rated in this question; the instructions for the test administrator are rated in question 3.2. The instructions or recommendations for the test taker form part of the test material and usually take up the first page or pages of the test booklet. The instructions must be standardised and put into standard Dutch. The instructions must include the following elements:
- Sample questions
- Information on how to record or note down the answers
- A strategy for guessing when the correct answer is unknown, or when two answers are equally likely or applicable
- Time limits

If indicated, the instructions must also include information on the anonymity of the test results. Where possible, practice items should be distributed to persons who have no experience with tests.

Notes for question 2.6: 'Are the items correctly formulated?'
Literature on test and questionnaire construction contains many rules for the formulation of items. Below is a non-exhaustive summary of rules that should be taken into account when rating the items (largely taken from Erkens & Moelands (1992), and Moelands, Noijons & Rem (1992)). If applicable, the rules for the tests below also apply to CBT tests (see question 2.14).

Open-ended questions
- Is the grammatical formulation of the item correct?
- Does the item contain overly complicated sentence constructions?
- Does the item contain unnecessarily difficult words?
- Does the item contain unnecessary insertions?
- Is the item expressed in an unnecessarily negative way?
- Can the formulation of the item give rise to misunderstandings?
- Is there any danger that a change of word stress in the item would clearly change its meaning?
- Does the item contain enough information for test takers to choose the right answer?
- Does the item give sufficient information on the desired length and structure of the answer?
- Does the test taker know whether a true/false answer must be accompanied with further explanation?
- Are the information and the problem definition clearly distinguishable?

Closed questions
- Is there possibly more than one correct answer?
- Are two things being asked about simultaneously?
- Does the item contain unclear passages?
- Do the alternatives contain unclear passages?
- Does the body of the question contain a clear question or task?
- Does the body of the question contain enough information for test takers to choose the right answer?
- Is it not possible to figure out from the item's characteristics what the right answer is?
- Is the body of the question free from superfluous information?
- Is the body of the question precise, concise and grammatically correct?
- Is the body of the question free of double negatives?
- If the body of the question includes a denial, is this made clearly visible?
- Are all distractors reasonably plausible?
- Does the correct alternative contain no repetition of terms from the body?
- Are the distractors free of words like 'always' or 'never'?
- Are no double negatives formed when the body is combined with one or more of the alternatives?
- Are the alternatives mutually exclusive?
- Does the content and grammar of the alternatives match properly with the body?
- Are the alternatives free of repetitions from the body or from each other?
- Are the alternatives logically ordered?
- Are the alternatives sufficiently distinguishable from one another?

Notes for question 2.7: 'What is the quality of the test materials?'
This exclusively concerns practical aspects which cannot be rated under one of the other questions for this criterion, such as:
- Is the text clearly legible?
- Is the test or questionnaire booklet well-organised? If other materials like blocks or tools are used, are these manageable and functional?
- Is the use of colours pleasant and functional? (see notes for question 2.15, fourth point)
- If applicable, are colours or symbols clearly distinguishable from one another, even for the colour-blind?
- Are the test materials sustainable?

Notes for question 2.8: 'Is the scoring system devised and explained in such a way that scoring errors can be avoided?'
This question calls for attention to points such as these:
- The scoring procedure must be clearly described.
- If scoring moulds are used, it must be explained how they should be laid on the answer sheets. The moulds must also fit properly on the answer sheets.
- If scoring moulds are used, they must indicate what version of the test they belong to. This is particularly important when the test is revised.
- It must be indicated how skipped items are to be scored.
- It must be indicated how many items can be skipped without causing the test to lose its value.
- If the test requires raters or observers, it must be indicated how to deal with differences between the raters and observers.

In order to prevent any possible scoring errors, a separate answer sheet should generally be preferred over multiple pages in a test booklet. Please note: For tests that are given with paper and pencil but scored by computer, the COTAN reviewer must be able to perform a score check (see question 2.10).

ADMINISTRATION BY COMPUTER

Notes for key question 2.9: 'Is the test standardised, or are decision rules for adaptive tests specified?'
Test administration via computer carries the same standardisation requirement concerning content and formulation of the items. With such tests, standardisation of the test time should receive extra attention. It is important that the time available for one item or the entire test should not depend on the system running the application.
Although the standardisation requirement (see question 2.1) is essentially the same for all tests, an exception is made for the item content and order of adaptive tests. The theory behind adaptive tests assumes that the test taker's skill level can be more efficiently assessed if the selection of items is constantly adjusted in response to the test taker's answers to previous items. For this type of test, the decision rules or algorithms used to create the test must be explicitly stated. How is the test started? How is the choice of the next item made? When does the test end? If either the starting procedure, selection procedure or ending procedure is not described, this question is rated as 'insufficient' (1). A rating of 'good' (3) can only be given if the choice of an algorithm is scientifically supported and the advantages and disadvantages of the choice are explained.

Notes for key question 2.10: 'Is there an automated or objective scoring system?'
A scoring system is called objective when the score values accorded to all possible answers by the participants are sufficiently agreed upon such that any qualified person who scores the items will, apart from clerical errors made during scoring, give exactly the same score. If the scoring is completely automatized, then the scoring is by definition objective. The rating in such a case is 'good' (3). However, this rating may only be given if the COTAN reviewer has access to enough information to check the accuracy of the scoring. In this case, 'scoring' refers to awarding a score to the items, adding up the item scores per test or scale, using item weighting if indicated, and converting these sum scores into norm scores with the norms table. This requirement for a check can mean that the test author may have to provide extra information not contained in the manual, such as keys, weights or norm tables.

Even if the test user has no access to such extra information, he should have access to information that enables him to interpret a test taker's results. This applies primarily to raw test or scale scores. If these are not listed in an automatically generated report and cannot be called up elsewhere in the application, the rating for this question cannot be higher than 'sufficient'.

If a few or all items on a test with open-ended questions are scored by hand, that test must be accompanied by model answers, scoring rules and rating instructions. These must clarify exactly what an answer must contain or what behaviour must be displayed in order to receive a particular score. For tests with closed items, only scoring rules are required. If this information

is not furnished, the rating for this question is 'insufficient' (1). In other cases, it is primarily the completeness and clarity of the extra material that will lead to a rating of 'sufficient' or 'good'.

Notes for key question 2.11: 'Are the test items free from racist, ethnically directed and sexist content, or any other content offensive to specific population groups?'
See the notes to question 2.3.

Notes for question 2.12: 'Has the software been designed in a way that prevents errors caused by incorrect use?'
Under no circumstances may the test results be negatively affected because a candidate has used the CBT software improperly. In addition to providing comprehensible instructions, there are many ways to prevent 'errors' stemming from improper use of CBT software. For this question, it is important that the test author has done enough to minimise the likelihood of errors caused by improper usage. Various precautions are important in this regard:

- Turning off unnecessary functions and hotkeys
- Cutting off access to the hard drive
- Making it impossible to start up other (unintended) software
- Making it difficult to close the CBT software prematurely, or without saving.

For tests taken via internet with the aid of a browser (Internet Explorer, Firefox, Safari, etc.) that presents the items and communicates the answers to the server, it is usually not possible to influence the client's computer. In that case, the test manual should state what precautions the test user must take.

The interface design also influences the chance of errors. This question does not call for a judgement on whether the user interface is properly designed, but the design may be taken into consideration when rating this question. If the user interface shows such features as an extreme number of navigational possibilities, nearly illegible texts or a confusing layout, a rating of 'insufficient' (1) must be awarded. If no serious problems surface when using the CBT software and it reacts as expected, a rating of 'sufficient' (2) should be awarded. A rating of 'good' (3) may only be awarded if it is genuinely difficult to start up any other (unintended) software, use unintended keys or key combinations, or leave the CBT software without saving.

When the test is administered by computer, either with a stand-alone, a network or an internet application, there is always the possibility that the test administration can be interrupted by a technical issue which is neither the fault of the test taker or the CBT software. In such cases it must be possible to restart, and (after the identification and possibly a repetition of the test instructions) the application must resume the test at the correct item, taking into account the remaining test time, if applicable. The COTAN reviewer is not expected to personally carry out an exhaustive check of the above matters. He must however judge whether the manual contains concrete information on the precautions that have been taken and the way in which these have undergone practical testing.

Notes for question 2.13: 'Are the instructions for the test taker complete and clear?'
Clear and complete instructions are important, as persons taking the test must not make 'errors' because they do not know how the CBT software works. The instructions must include the following elements:

- Sample questions
- The way the software works (including the method of answering)
- A strategy for guessing when the correct answer is unknown, or when two answers are equally likely or applicable
- The available time per test or per item

Also important:

- For adaptive tests, the procedure of adaptive testing must be explained
- If applicable, the instructions must also include information on the anonymity of the test results
- For persons who have no experience with the type of test in question, practice questions should be distributed that the client must answer and should receive feedback on.

Unclear or incomplete instructions or too extensive instructions (for example when instructions are given on how to answer every individual item) lead to a rating of 'insufficient' (1) for this question. The rating of 'good' (3) may only be awarded if it is possible to consult the instructions while taking the test.

Notes for question 2.14: 'Are the items correctly formulated?'
See the notes to question 2.6. It is also important to point out that for tests taken by computer, and especially for adaptive tests, the COTAN reviewer must be able to view all items. This can mean that the test author will have to supply a summary of all items, for review purposes only.

Notes for question 2.15: 'What is the quality of the design for the user interface?'
Below, aspects are mentioned requiring attention when rating the user interface. These aspects must be rated for the recommended standard installation and computer setup. A negative rating on one of these aspects can be enough to produce a rating of 'insufficient' (1), if they seriously impede the benefits of using the instrument.

- Is the screen design consistent? This concerns the following features of screen design:
  - Symbols must always have the same function
  - Colours must be used consistently and must always have the same function
  - Information such as items, instructions, answer fields and so forth must always appear at the same location, or distinctions between types of information must always be made in the same manner
  - Fonts and letter sizes must be used consistently.
- Are items effectively arranged on the screen? The effective arrangement of the display depends on various factors:
  - Are the different types of information (instructions, items, answer fields) clearly distinguishable from one another?
  - Are the buttons clearly recognisable and is the function of the buttons always clear? For example, what does the button <close> actually close: the test itself or only the instructions?
  - Can the items and instructions be read without having to scroll?
  - Are certain types of information, such as the instructions, easy to find?
  - Is it always clear for test takers where they are in the application, and what actions they must perform to reach a desired location?

To achieve this, one must check whether it is obvious enough that a person without any computer experience is still able to take the test without inducing test bias.

- Is the on-screen information legible? Legibility is improved when:
  - No more than two fonts are used
  - No more than three font sizes are used
  - No words are in italics
  - Words are not underlined except in the case of a hyperlink.
- Is the colour use pleasant and functional? It is important that colour be used in a way that promotes the organisation and legibility of the on-screen images. Functional colour use means that colours have a certain meaning or that the screen is easier to read, for example by giving certain items or the answer field a different colour. It is certainly undesirable to use a large number of colours or employ colours for no apparent reason. 'Pleasant' colour use refers to the choice of particular colour combinations, or the contrast between hues. For example, certain colour combinations, as well as poorly contrasting colours, are difficult to distinguish. When using colours, one must also not forget that the test should be generally suitable for colour-blind people, and that the colour use must not pose any disadvantages to this group.
- Are the visual and auditory materials functional? In this context, 'visual materials' refers to any possible visual materials such as animations, film clips and static illustrations. It is important that both visual materials and sound clips have a clear function, and that they are not introduced merely to 'beautify' the CBT software. Another remark here is that the functionality of the visual and auditory materials suffers if they are difficult to read or understand.

Notes for question 2.16 'Is the test sufficiently secured?'
A test is 'well' secured if every effort has been made to protect the access to the test, test materials and test results.

- Security measures for test access are needed to ensure that the person taking the test is indeed the person who is supposed to do this. Some form of identification is therefore necessary. This might involve the use of passwords and user names, compulsory identification with an identity card or driver's licence, or the use of webcams.
- Security of the test materials is important first of all because, to preserve validity, test takers must not be in a position to copy items, information on the algorithms or scoring rules to another computer or printer. Second, it is important that no information about the items can be easily obtained. Thus, if all the items are contained in an item bank, only authorised persons should be able to gain access to it. In adaptive tests, items may also start to become familiar because one item might be included much more often in the test than some other item. In some cases, the test author should incorporate a mechanism like the Sympson-Hetter method (1985; see also Stocking & Swanson, 1993), which checks for possible over- or under-use of the items.
- Security measures for the test results are needed to prevent cheating, such as unauthorised changes in the results, and to sufficiently guarantee the privacy and anonymity of the test taker.

For a rating of 'good' (3), the manual must contain concrete material demonstrating that all three aspects of security have been sufficiently addressed. The rating is 'insufficient' (1) when no such information is given, or the information supplied reveals that one or more of these aspects have not been addressed. The rating of 'sufficient' (2) is given when attention is paid to all three aspects, but technical and procedural improvements could still be made.

Rules for determining final rating for criterion 2 Quality of the test materials Paper-and-pencil version		
All three key questions are rated '3'.	Sum score 2.4 through 2.8 ≥ 11	good
	Sum score 2.4 through 2.8 = 9 or 10	sufficient
	Sum score 2.4 through 2.8 ≤ 8	insufficient
Key question 2.2* and/or 2.3 is rated '2' and the other key questions are not rated '1'.	Sum score 2.4 through 2.8 ≥ 11	sufficient
	Sum score 2.4 through 2.8 ≤ 10**	insufficient
At least one of the three key questions is rated '1'.		insufficient
* For key question 2.2, both sub-questions can be applicable; in that case, the lowest rating is chosen. ** If questions 2.4 through 2.8 are all rated '2', the final rating is 'sufficient'.		

Rules for determining final rating for criterion 2 Quality of the test materials Computer version		
All three key questions are rated '3'.	Sum score 2.12 through 2.16 ≥ 11	good
	Sum score 2.12 through 2.16 = 9 or 10	sufficient
	Sum score 2.12 through 2.16 ≤ 8	insufficient
Key question 2.10 and/or 2.11 is rated '2' and the other key questions are not rated '1'.	Sum score 2.12 through 2.16 ≥ 11	sufficient
	Sum score 2.12 through 2.16 ≤ 10*	insufficient
At least one of the three key questions is rated '1'.		insufficient
* If questions 2.12 through 2.16 are all rated '2', the final rating is 'sufficient'.		

3 Quality of the test manual

This criterion examines the comprehensiveness of the information offered by the manual to the user. On the one hand, this concerns practical indications for administration, scoring and interpretation (sometimes provided in a separate user's guide), and on the other hand information on research that has been performed with the test (sometimes provided in a separate technical manual). The user requires both types of information to reach a decision on what conclusions can be drawn from a test score. This information must be clearly set out for the user and available either in paper or digital form. For tests administered by computer, specific directions must be given for the installation, the starting procedure and the use of the test (sometimes provided in a separate installation manual); see questions 3.8 through 3.10.

Notes for key question 3.1: 'Is a test manual available?'
Every test must be furnished with a test manual. Dissertations or collections of research papers are not regarded as a manual.

Notes for question 3.2: 'Are the instructions for the test administrator complete and clear?'
The main objective of the notes for the test administrator in the manual is to ensure the standardisation of the test. The manual should describe as explicitly as possible what the test administrator must or must not say (the suggestion that 'the test administrator explains the purpose of the test' is insufficient), and the tasks the administrator must perform (such as arranging the test materials in a certain order for an ability test). The test manual must also state how the administrator should deal with questions: for example, it might state standard answers to the most common questions. The manual must indicate the degree of support that may be given and the aids that the test taker may use. For a computer-administered test, the manual must state what computer skills the test taker has to possess, and the circumstances under which the test should be administered (comfort, work space, lighting and so on).

Questions for criterion 3 Quality of the test manual				
		ins.	suf.	good
Key question 3.1	Is a test manual available? If the rating of this question is negative (1), skip the other questions for this criterion and continue with criterion 4.	1		3
3.2	Are the instructions for the test administrator complete and clear?	1	2	3
3.3	Is it specified how the test can be used and what the limitations of the test are?	1	2	3
3.4	Is a summary of the research results presented in the manual?	1	2	3
3.5	Are case descriptions used to explain how to interpret the test scores?	1	2	3
3.6	Is it indicated what kind of information may be important for the interpretation of the test scores?	1	2	3
3.7	Is it specified what professional qualifications are required to administer and interpret the test?	1	2	3
Extra questions for administration by computer				
3.8	Is information supplied on the installation of the computer software?	1	2	3
3.9	Is information supplied on the operation and capabilities of the software?	1	2	3
3.10	Are there enough options for technical support?	1	2	3

Notes for question 3.3: 'Is it specified how the test can be used and what the limitations of the test are?'
A manual must be complete, accurate, and clear about the applicability and the limitations of the test. It must therefore be clear to the test user what constructs the test is meant to measure, what target group the test is meant for and what the test's function is (classification, selection and so forth). The test limitations must also be described. In this regard, various suggestions might be made for using the test, depending on the specific situation the test is intended for. For example, is it stated that decisions on educational classification should not be taken on the basis of a single test score? Has the relationship between the test score and the subsequent learning process been specified in cases of educational progress? Can test results obtained in a clinical situation lead to empirically founded conclusions or can they serve only as working hypotheses? Has it been pointed out that test scores alone should not be used as a basis for decisions relating to vocational guidance? Is stated for which job types tests for personnel selection are intended and what the critical content is of these jobs?

Notes for question 3.4: 'Is a summary of the research results presented in the manual?'
For both prospective test users and COTAN reviewers, the manual will be the principal source of information. Users often lack easy access to dissertations, articles in foreign journals, research reports or other published materials, and the technical details are not easy to understand. The manual must therefore provide a summary of norming, reliability and validity studies. This should be informative and thorough enough that potential users can judge whether a test is suitable for their purposes and is of the required quality. A COTAN reviewer may sometimes want to consult the original literature, so the manual must contain references to it. If applicable, a summary of the design and results of the calibration and simulation study should also be included. If new research provides useful additional information, users should be informed by means of supplements to or revisions of the manual. The internet provides excellent opportunities for distributing handy addenda to users.

In this question, only the availability of the information in the manual is assessed. The quality of the research designs and results are not discussed here, because they are evaluated in the criteria 4, 5, 6 and 7.

For so-called research instruments, there is often no manual. In such cases, the question receives a negative rating, but original articles, dissertations, and reports will be taken into account for the evaluation of other criteria.

Notes for question 3.5: 'Are case descriptions used to explain how to interpret the test scores?'
A manual must include several case descriptions and sample reports.

Notes for question 3.6: 'Is it indicated what kind of information may be important for the interpretation of the test scores?'
Is it explained what other variables contribute to the prediction? Is there a discussion of how background variables and test experience might influence the scores?

Notes for question 3.7: 'Is it specified what professional qualifications are required to administer and interpret the test?'
In the manual attention must be paid to the professional qualifications of the intended test users. For example it could be described which kind of professionals are suited to administer and interpret the test, based on their education and/or work experience. An appropriate description should also be given of the knowledge and skills considered necessary for administration and interpretation of the test.

EXTRA QUESTIONS FOR ADMINISTRATION BY COMPUTER

Notes for question 3.8: 'Is information supplied on the installation of the computer software?'
Information on the necessary hardware and software and the way to install the CBT software is a requirement. As for the hardware, it is important to mention the requirements for CPU, minimum memory space, hard disk space, monitor and video card, input devices and exchange devices such as CD-ROM players. Information on the required network card or sound card might also be required. As for the software, it is important to mention the operating system under which the test will function and any other required software, such as browsers or particular plugins. The method for installing the software should be described step by step and supported by screenshots wherever possible.

In the absence of a description of the required hardware or software, or the installation of the CBT software, the rating is 'insufficient'. The description of the CBT software installation can be considered present if the CBT software automatically installs itself. Only with an extensive description of the required hardware and software, as well as a proper description of the CBT software installation (apart from self-installing systems), can this question receive a rating of 'good'.

Notes for question 3.9: 'Is information supplied on the operation and capabilities of the software?'
For every CBT, information must be given on the operation and capabilities of the software, such as choice of settings, the possibility of group summaries, and analysis and reporting options. If information on any of these aspects is insufficient or completely lacking, the rating is 'insufficient'. In other cases, the clarity and completeness of the information given is decisive for the rating of 'sufficient' or 'good'.

Notes for question 3.10: 'Are there enough options for technical support?'
If the test user has questions about the CBT software, or when it malfunctions, technical support must be available. This can take the form either of documentation on common problems, such as a section on 'FAQs' or of a help desk whose availability and accessibility is clearly indicated in the manual.

This question can be rated 'good' only when written or digital documentation on problem-solving, as well as the option of consulting a help desk, are present. Only when there is no documentation on problem solving and the test user cannot fall back on a help desk, is the rating 'insufficient'. In all other cases, the rating is 'sufficient'.

Rules for determining final rating for criterion 3 Quality of the test manual Paper-and-pencil version		
The key question is rated '3'.	Sum score 3.2 through 3.7 ≥ 13	good
	Sum score 3.2 through 3.7 = 11 or 12	sufficient
	Sum score 3.2 through 3.7 ≤ 10	insufficient
The key question is rated '1'.		insufficient
Rules for determining final rating for criterion 3 Quality of the test manual Administration by computer		
The key question is rated '3'.	Sum score 3.2 through 3.10 ≥ 19	good
	Sum score 3.2 through 3.10 = 17 or 18	sufficient
	Sum score 3.2 through 3.10 ≤ 16	insufficient
The key question is rated '1'.		insufficient

4 Norms

Scoring a test results in a so-called raw score. Raw scores are determined by various properties of a test, such as the number of items, the choice of the time limit, the degree of difficulty, the popularity of the items and the circumstances under which the test is administered. This means that raw scores are difficult to interpret. In general, the raw score can only be understood by referring to a norm.

There are two types of norm scores (APA, 1999). With the first type, the raw score earned is compared with that of other test takers. This type of interpretation is known as norm-referenced interpretation. The scores are compared with the distribution of scores from a reference group. The goal is to determine how a test taker's score compares with the scores of other individuals with whom a useful comparison can be made (on the basis of similarities in age, grade, job, etc.). This type of norms is also referred to as relative norms.

With the second type, results are not compared with those of others, but are interpreted absolutely: in other words, the results are compared with an absolute norm. This type of interpretation is known as content- or criterion-referenced interpretation. In this type of norming, certain standards or cut-off scores are set. In content-referenced interpretation, these standards are set in one way or another by experts or raters. The norm may be directly derived from a description of the domain of skills or subject matter that one must have mastered. This type of norms is also known as absolute norms. With criterion-referenced interpretation, cut-off scores are derived from research data. This type of norming requires not only test data but data on the criterion as well. When one single cut-off score is used, it can mark the difference between failing and passing, or rejection and admission. When there is more than one cut-off score, it is possible to distinguish different skill levels from one another.

If no norms are provided, then the final rating on this criterion is in principle 'insufficient'. However, well-argued exceptions can occur, for example with tests in which purely intra-individual comparison is indicated and justified, as with ipsative tests or tests that measure progress in time. In such cases, this criterion may be labelled as 'not applicable'.

Questions for criterion 4 Norms General key questions				
		ins.	suf.	good
Key question 4.1	Are norms provided? If the rating of this question is negative (1), skip the other questions for this criterion and continue with criterion 5.	1	Not applicable	3
Key question 4.2	Are the norms up to date? If the rating of this question is negative (1), skip the other questions for this criterion and continue with criterion 5.	1 1	2 2	3 3

In cases of norm-referenced interpretation, proceed with question 4.3.
In cases of content-referenced interpretation, proceed with question 4.8.
In cases of criterion-referenced interpretation, proceed with question 4.11.

Questions for criterion 4 Norms Norm-referenced interpretation					
		ins.	suf.	good	
Key question 4.3	What is the quality of the norm groups provided? a. Are the norm groups large enough? b. Are the norm groups representative? If the rating for question 4.3.a or 4.3.b is negative (1), skip the other questions for this criterion and continue with criterion 5.	 1 1	 2 2	 3 3	
4.4	Are the meaning and the limitations of the norm scale made clear to the user, and is the type of norm scale in agreement with the purpose of the test?	1	2	3	
4.5	Is there information about the means, the standard deviations, and the score distributions?	1	2	3	
4.6	Is there information about possible differences between subgroups (for instance with respect to gender and ethnicity)?	1	2	3	
4.7	Is there information about the accuracy of the measurement, and the corresponding confidence intervals? a. standard error of measurement b. standard error of estimate c. test information function / standard error	 1 1 1	 2 2 2	 3 3 3	 Not applicable Not applicable Not applicable
Content-referenced interpretation					
If the cut-off scores are determined with the aid of raters: what is the quality of the standard procedure for determining them?		ins.	suf.	good	
Key question 4.8	Is there sufficient agreement between the raters?	1	2	3	Not applicable
4.9	Are the procedures for determining the cut-off scores correct?	1	2	3	Not applicable
4.10	Have the raters been properly selected and trained?	1	2	3	Not applicable
Criterion-referenced interpretation					
If the cut-off scores are empirically supported: what is the outcome and quality of this research?		ins.	suf.	good	
Key question 4.11	Do the research results justify the use of cut-off scores?	1	2	3	Not applicable
4.12	Is the research sample appropriate for the intended use?	1	2	3	Not applicable
4.13	Is the research sample large enough?	1	2	3	Not applicable

Notes for key question 4.1: 'Are norms provided?'

Norms, whether they are intended for norm-referenced interpretation (like norm tables) or content- and criterion-referenced interpretation (such as cut-off scores or expectancy tables), must be available at the time the test is published for actual use. The following situations can cause this question to be answered in the negative:

- The data mentioned above are not provided: for example, only means and standard deviations are given for the norm group or research sample.
- For tests intended for interpretation at group level, norm tables are provided that are based on individual scores, and vice versa (see also 4.3.a).
- After the norms were collected, alterations were made in the test itself, for example in the items or the instructions.
- The norms were collected with the use of paper and pencil, while the version being rated is a computer version, or vice versa. For questionnaires, this generally has little influence on the applicability of the norms (Bartram, 2005; King & Miles, 1995; Mead & Drasgow, 1993). For cognition and skills tests and tests for which a time limit is in force, however, new norms will have to be collected.

Notes for key question 4.2: 'Are the norms up to date?'

Norms are susceptible to deterioration. Of all the psychometric properties of a test, norms are the most sensitive to such factors as social changes, educational changes and changes in functions. Consequently, a test either has to be renormed from time to time, or the test author has to perform research to show that there is no need for renorming. For example, with intelligence tests one must take into account the Flynn effect, which causes norms to become outdated (see Resing & Drenth, 2007, pp. 142-145). This effect is estimated at 3 IQ points for every 10 years, or 4.5 IQ points for every 15 years. This is equivalent to approximately one standard error of measurement (with a reliability of .91). Such an effect probably also holds for related tests such as test batteries for general or specific cognitive capacities. Nothing is known about effects of this type for personality tests.

Comparison of data in the manuals of several Dutch tests has yielded the following information. For the *Amsterdamse Beroepen Interessen Vragenlijst (Amsterdam Vocational Interest Questionnaire)* (Evers, 1979, 1992) differences were found over a period of 16 years that rose as high as two standard deviations. Over a period of more than 20 years, the *Nederlandse PersoonlijkheidsVragenlijst (Dutch Personality Questionnaire)* found maximum differences of 1.4 standard deviation for the selection norm group, 1.2 standard deviation for the general norm group and 0.5 standard deviation for the norm group of psychiatric patients (Luteijn, Starren & van Dijk, 1985; Barelds, Luteijn, van Dijk & Starren, 2007). It should be noted here that items were changed in both these questionnaires. American research (Twenge, 2000) shows that anxiety and neuroticism scores in the US have increased over the past 40 years by approximately one standard deviation.

The German review system for test quality (Kersting, 2006) recommends a period of eight years for renorming, but does not call for strict enforcement. The APA standards (APA, 1999, p. 59, Standard 4.18) state that "... so long as the test remains in print, it is the publisher's responsibility to assure that the test is renormed with sufficient frequency to permit continued accurate and appropriate score interpretations". The APA specifies no duration in this area. On the basis of the above findings, and in an attempt to balance what is achievable with what is desirable, the COTAN has arrived at the following rule. To draw the user's attention to possibly outdated norms, the qualification 'the norms are outdated' will be added to the review of tests for which no new renorming or calibration studies have been performed in the last 15 years. After five more years without renorming research, this qualification will be changed to: 'Norms unusable because they are outdated' and the rating 'insufficient' will be given. Once per year, all test descriptions in the online *Documentation of Tests and Test Research* will be updated to reflect this. To evaluate the degree to which norms may be outdated, it is important to state the year or period of the data collection. If this is not stated, the rating for norms will be 'insufficient'.

NORM-REFERENCED INTERPRETATION

Notes for key question 4.3: 'What is the quality of the norm groups provided?'

Basically, norms must be presented for all purposes for which the test author recommends the test (see question 1.1). It may turn out that the groups for which norms are presented only partly cover the intended applications. For instance, when a test author indicates that a test is intended for both vocational guidance within technical schools and for selection for technical jobs at this level, norms should be provided for both situations. However, it would not be realistic to require norms for every technical job.

A norm group has to meet two requirements to fulfil its goal, namely to supply a reliable set of reference points. First, it must be sufficiently large, and at the same time it must be representative of the intended group. Notes are given below for the rating of both these aspects. The rating for question 4.3 can only be 'good' if both these aspects (questions 4.3.a and 4.3.b) are rated as 'good'. The rating is 'insufficient' when at least one of these questions is rated as 'insufficient'. In all other cases, the rating is 'sufficient'.

Notes for question 4.3.a: 'Are the norm groups large enough?'

Recommendations on the desired sample size are scarce in the literature (Angoff, 1971; Campbell, 1971). These recommendations are either based on the calculation of standard errors in parameters such as the mean and the median, or on experiential data with respect to the stability of scale values. A combination of these two, linked to the importance of the decisions to be made on the basis of the test, has resulted in the following rules for rating:

Tests for important* decisions at individual level (for instance, personnel selection, placement in special educational programmes, admission for/discontinuation of clinical treatment, certification).	N ≥ 400 300 ≤ N < 400 N < 300	good sufficient insufficient
Tests for relatively less important decisions at individual level (for instance, evaluating educational progress and general descriptive use such as vocational guidance and admission for therapy).	N ≥ 300 200 ≤ N < 300 N < 200	good sufficient insufficient

* Important decisions are understood as decisions taken on the basis of the test scores that are essentially, or in the short term, irreversible, and on which the test taker has little influence.

The requirement with respect to sample size naturally applies to each *norm group* for which norming takes place. In cases such as developmental tests which are normed for various age groups, this may cause confusion. If norming is performed separately for different age groups, grades or types of schools, the sample size of each norm group is a significant factor. However, if continuous norming or fit procedures are applied (using the data of all age groups concurrently) the sample size of the individual age groups can be smaller, because this procedure yields more efficient estimators than classical norming.

Bechger, Hemker, and Maris (2009) performed a study comparing the equivalence of the group sizes in classical and continuous norming. This study shows that, in a special case, the required group size for continuous norming can be lower. Here, they use the standard error of the mean as a parameter which can be used as an indicator of the accuracy of the norms. The guiding principle for determining group size with continuous norming is that the accuracy of the norms must at least equal the level of that obtained with classical norming; in other words, norming in which norm data are calculated separately for each group.

Two side remarks need to be made both on this study and on the guidelines the COTAN has derived from it. Firstly, a number of statistical assumptions are made when doing the calculations, such as the assumption that the variations in the subgroups are equal, that the scores within each subgroup are normally distributed and that the regression of the test score on age is linear. Failure to satisfy these assumptions can lead to larger standard errors and thus to a larger number of required observations than mentioned by Bechger et al. (2009). There are also different variants of continuous norming, but in the study mentioned above only one variant with eight groups was completely calculated.

Secondly, in the case of continuous norming, it is not necessary to have the same number of people in each subgroup to achieve equal precision for all subgroups. With continuous norming, the groups in the middle require fewer observations than the extreme groups. For the sake of transparency, the COTAN's guidelines insist on the same number of persons per subgroup. The consequence of this is that the numbers chosen for the extreme groups cause a small loss in accuracy of measurement, but the accuracy in the middle groups is far greater than that achieved for these groups by classical norming. The guidelines below must therefore be viewed as the lower limit of what is theoretically desirable.

Guidelines for subgroup size in continuous norming with eight subgroups		
Tests for important* decisions at individual level (for instance, personnel selection, placement in special educational programmes, admission to/discharge from clinical treatment, certification).	N ≥ 150 100 ≤ N < 150 N < 100	good sufficient insufficient
Tests for relatively less important decisions at individual level (for instance, evaluating progress and general descriptive use of test scores in such areas as vocational guidance and admission to therapy).	N ≥ 100 70 ≤ N < 100 N < 70	good sufficient insufficient
* Important decisions are understood as: decisions taken on the basis of the test scores that are essentially, or in the short term, irreversible, and on which the test taker has little influence.		

For a comprehensive explanation of the way in which these guidelines were created, see the study by Bechger et al. (2009). Only a brief explanation is given here. For the standard error of the mean, values are calculated belonging to the sample sizes that serve as cut-offs for the qualifications 'insufficient', 'sufficient' and 'good' for norming of separate groups, namely 400, 300 and 200 persons. For a standard deviation of 15 (which is customary in intelligence tests), the standard errors of the mean are respectively 0.75, 0.87 and 1.06. Based on the regression approach of continuous norming, it is then possible to calculate the standard error of the mean for each subgroup; this holds for any number of groups and any sample size. For eight subgroups, such as the eight grades in primary school, and with a size of 100 persons per group, the error of the mean is equivalent to about .54 for groups 4 and 5, .63 for groups 3 and 6, .77 for groups 2 and 7 and .96 for groups 1 and 8. With classical norming and a group size of 300 persons, the standard error equals .87 in all groups. When continuous norming is used for eight groups containing 100 persons each, the accuracy is therefore improved in six of the groups and is worse in two groups than with classical norming. Although this last statistic is of course undesirable, the degree of worsening is limited and the gains in the middle groups are large. The COTAN has therefore declared the guideline of 300 persons with classical norming to be equivalent to that of 100 persons with continuous norming, when there are at least eight groups. The group sizes of 150 and 70 persons originated in a similar fashion. The accuracy worsens slightly only in the two extreme groups, but is better in the other six groups, in comparison with 400 and 200 persons respectively when using classical norming.

The above guidelines are specifically meant as examples. Their primary limitation is that the guidelines apply only to the situation in which eight subgroups can be distinguished. The second limitation is that the example applies only to the standard error of the mean and that other moments of the distribution are not included in the example. The third limitation is that the example rests on the idea that the statistical assumptions are satisfied. When continuous norming is used, it is thus the author's duty to indicate what sample size in classical norming is equivalent to the sample size being used in his test. Notice that it is not enough to simply show that the standard error of the mean is equivalent; equivalence must also be shown for other moments of the distribution, such as the standard deviation. If assumptions are being used to determine that equivalence, it must be indicated whether those assumptions have been satisfied.

When a test is intended purely to make statements at group level, other requirements for the sample size apply, because the standard error of the group mean is generally much smaller than that for individual scores. For instance, Angoff (1971) claims that the distribution of individual scores for tests of school performance is 2 to 2.5 times greater than the distribution of the group mean. A combination of this principle with the rule for individual norms leads to the rating rules given in the table below.

Tests for research at group level.	K ≥ 40 30 ≤ K < 40 K < 30	good sufficient insufficient

Here, K is the number of groups, and each group must consist of at least 25 persons. Examples of this type of tests are questionnaires on school conditions, job satisfaction, job conditions and organisational culture. For representativeness and supplementary statistical information, the same requirements apply, mutatis mutandis, as formulated in questions 4.3.b through 4.7, unless otherwise indicated.

Notes for question 4.3.b: 'Are the norm groups representative?'

A sample is representative if its composition corresponds for a number of variables to that of the population in question, and the sample is obtained with the use of a random sampling model. In a random sampling model, every element of the population has an equal chance of being included in the sample. To judge whether the norm groups are representative, both an adequate definition of the population and the sampling design or data collection process must be provided. This requirement holds for every group for which norms are provided, with both classical and continuous norming. It is frequently the case that the information offered is so limited that one cannot even tell what the target population is. It can sometimes be unclear whether a test is intended for national, regional or local use. Alternatively, the data collection could either cover a cross-section of the population or it could target individuals with certain characteristics (for instance, only people who require psychological assistance, or people with a specific educational background).

In any case, the composition of a sample must be described in relation to the variables of age, gender, ethnicity and region, as experience shows that these variables lead to score differences between subgroups in a wide range of tests and questionnaires. Depending on the constructs the test is supposed to measure, it may be wise to describe the composition in relation to variables such as degree of urbanisation, socioeconomic status, educational level, reason for test use, job level, line of business, and absence or presence of referral or clinical diagnosis. To establish whether the distributions in the samples being described are in agreement with those in the corresponding populations, a description of the populations in question must always be supplied. For more general background variables, one can generally turn to the data supplied by *Statistics Netherlands* (CBS). If the variables being used in the sampling model show shortages at certain levels, this may be corrected to a limited degree by weighting. In case of underrepresentation, a maximum factor of 2 is acceptable.

When registering ethnicity, it can be a problem that the data are not available or that registration was not permitted. There are also varying definitions of ethnicity which are not always decisive, or sufficient for research purposes. Finally, the composition of the target population in this regard is not always known. This does not relieve the test author of the obligation to try to supply the most complete information possible. Because of these considerations, norm groups can only receive a rating of 'good' when they originate from a random sampling model that strives for national representativeness. Two very common ways of collecting data do not meet this requirement: regional norms and samples of convenience. Notes are given below for the rating of both these sample types. Norms based on samples such as these can only receive a maximum rating of 'sufficient'.

Regional norms

When the test is intended for national use, nationally distributed norms must be collected, because the scores on many types of tests display regional differences. Regions exert their influence primarily because regions are associated with variables that cause these scoring differences, such as socioeconomic status, educational level and ethnicity. If the author can show that the principal background variables which make up part of a regional sample correspond with the national population, the highest possible rating on question 4.3.b for a regional sample is 2. Which background variables are important here depends on the expected correlation between background variable and test score. With a test for language skills, it can be expected that the scores will be linked with ethnicity. If a regional sample is based on a sample from a large city in the western part of the country, non-native respondents will probably be over-represented with respect to the rest of the Netherlands, meaning that the mean test score and the norms will not be representative of the Netherlands as a whole. In this case, data must be presented on ethnicity within both the sample and the population, and a correction must be made, if needed, for the unrepresentative composition of the sample. Consequently, a score of 2 on question 4.3.b means that the maximum possible rating for regional norms is 'sufficient'. The normal procedures for weighting then dictate that the choice for a rating of 'sufficient' is dependent on the rating for questions 4.4 through 4.7. Important variables in this context are gender, age and ethnicity, but the purpose of psychological testing (for example, selection or career counselling) may also be important.

Samples of convenience

Data collection quite frequently makes use of samples of convenience such as students who require assistance for their choice of school, psychology students who happen to be available, or the clients of a vocational guidance service. In general, samples of this kind make poor norm groups because they have not been controlled for variables that are related to the test score. Besides, they cannot be considered representative for the intended populations, such as middle school students, students in higher education, or all Dutch employees who have one certain type of job or are of a certain level.

Samples of convenience are actually not samples in the strictest sense of the word: this usually refers to the entire client population that fills in a questionnaire or a test in a certain period or on a certain occasion. There is no guarantee that each member of the target population has an equal chance of being included in the sample. The data collection is not based on a sampling model. The problem with samples of convenience is that one does not precisely know what one is collecting. For example, if a sample like this is based on the clients of a number of vocational guidance agencies, can a group like this be considered a true cross-section of the Dutch population with the same age and education, or is there reason to suppose that people who have difficulties in choosing an occupation are different from the rest of the Dutch population? Or can a

norm group comprised of the applicants tested by personnel selection firm X with three offices throughout the country be used for the applicants of firm Y? Norm groups of this type will generally be rated 'insufficient', as their composition is either unknown or not subject to control. Some test authors do consider the size of the norm group to be an argument for the representativeness, validity or applicability of the norms. The size of such norm groups is indeed usually not a problem (numbers in the thousands are not an exception), but the size of a sample in itself reveals nothing about its representativeness, nor about its usability. For example, the clients of firm X consist primarily of companies in the ICT sector. The norms based on this group of applicants may well be suitable for other personnel selection firms who specialise in ICT, but not for firms whose clients come primarily from other occupational fields.

Norms based on samples of convenience can be seen as a special variety of regional norms, and regional norms can in turn often be a sort of sample of convenience. For instance, one might consider the clients of firm X as a sample of all applicants in the Netherlands. Just as with regional norms, the quality of a sample of convenience can be rated 'sufficient' if an exhaustive description of variables that are possibly relevant to the norm group is given. However, it is not enough to demonstrate representativeness of the intended target group in terms of gender, age and ethnicity alone; there must also be variables that are associated with the purpose of the test, such as the line of business and the job in selection situations, and the type of disorder in clinical situations. The normal procedures for weighting dictate that the rating of 'sufficient' is dependent on the rating for questions 4.4 through 4.7. The motivation for a rating of 'sufficient' in these cases is that such norms can be successfully used when the user knows to which group a client or applicant can be compared.

Notes for question 4.4: 'Are the meaning and the limitations of the norm scale made clear to the user, and is the type of norm scale in agreement with the purpose of the test?'
For the conversion of raw scores into derived scores, one can make a choice from three types of norms (Drenth & Sijtsma, 2006): proportional norms, norms based on ranking, and norms based on means and standard deviation.

One well-known example of proportional norms is the old-fashioned Intelligence Quotient (IQ), in which mental age is divided by chronological age. Currently, this method of IQ calculation is scarcely in use, if at all. Another example is the Didactic Age Equivalent (DAE), which evaluates the level of a child's learning performance in terms of the duration of schooling (in months) needed to reasonably expect this performance in proportion to the actual duration of schooling. There are many practical and theoretical objections to proportional norms. For an extended discussion, see Evers and Resing (2007). The COTAN rejects the use of proportional norms in general and of DAEs in particular. Consequently, tests reported solely in terms of DAEs are rated as 'insufficient' for norms. The COTAN would prefer that DAEs completely disappear, but some tests report DAEs alongside standard scores or ranking scores. In such reports, the standard scores and ranking scores must definitely take precedence and be accompanied by an extensive explanation of how these systems are to be used. Additionally, the manual must contain an explicit warning about the limitations of DAEs.

Norms based on ranking are percentiles and scale systems derived from them, such as vigintiles, deciles and the A to E system used by Cito. Examples of norm types based on means and standard deviation, here referred to as deviation norms, include stanines, C-scores, T-scores and deviation IQs. Within deviation norms, a further distinction can be made between linear transformations and normalised transformations. In principle, normalising transformations should be used, unless the raw scores already approximate a normal distribution. If more than one norm group is being distinguished, as with successive age groups or school grades, it is better to begin by applying fit procedures to the cumulative distribution (Laros & Tellegen, 1991), because transformations based on the direct conversion of the cumulative proportions observed are very sensitive to sampling fluctuations.

When age or grade norms are provided, an excessively wide age or grade interval can have the effect that performance at the beginning of the interval is overestimated and underestimated at the end. Ability tests for young children are particularly susceptible to this effect, with a variation of 10 or more IQ points within a one-year period. Even in the first years of secondary education, the difference between two consecutive years may amount to as much as half a standard deviation. This kind of bias can be easily prevented by expanding the number of norm tables for age and grade norms, by using continuous norming or by correcting for the day on which the test is administered (so-called instruction day correction), which in fact comes down to the same thing. If a test is intended only for use during a particular period of the school year (as with some Cito tests), this must be clearly indicated, and the norm data must have been collected in the corresponding period. Age and grade norms must state the time of year in which the norms were collected.

The various scale types found in both ranking and deviation norms differ in the number of score units that are employed. Systems with many units, such as percentiles and deviation IQs, make more precise distinctions possible than systems with few units, such as A-E scores and stanines. The choice of a particular system depends on the purpose and the features of the test. If the aim of the test is to establish broad differentiation between persons, one should choose a precise system with many units, but only under the condition that the range of possible raw scores should also offer enough possibilities for differentiation. For example, it is pointless to use percentiles (which have 100 units) when the minimum raw score on the test is 0 and the maximum is 20, so a total of only 21 score units. Opting for a rough system is detrimental to the differentiation, but may make the outcomes more readily comprehensible. A choice of this type is preferable when only a broad indication is required. No matter what scale system is used, the features and the possible pros and cons of the system should be described and the reasons for choosing the particular scale should be given.

Regarding tests intended for research at group level, no norm tables are usually provided: a statement of the mean and the standard deviation of the norm group or groups will suffice.

Notes for question 4.5: 'Is there information about the means, the standard deviations, and the score distributions?'
Data on the means, the standard deviations, and the score distributions must be provided for each group. Aspects of the distribution such as kurtosis, skewness and bimodality are relevant, as well as possible differences in these parameters between norm groups. For example, it may be the case that the scores on a questionnaire are more or less normally distributed in one group, while 50% of the participants in another group obtain the lowest score (the so-called bottom effect). Another example might be the bottom or ceiling effects in tests for cognitive abilities shown by groups with a low or high level of education, respectively. This causes the test to discriminate less well in these groups.

Notes for question 4.6: 'Is there information about possible differences between subgroups (for instance with respect to gender and ethnicity)?'
The data mentioned in question 4.5 must also be stated for possible subgroups. There are various reasons why subgroup differences must be studied and reported:

- The results may show adverse impact.
- This can constitute an extra reason for performing research on both test and item bias.
- Supplying these data enables test users to make their own decisions about whether to include them in their interpretation. Even when there appear to be significant differences between subgroups, it is not always desirable to employ norm tables for each subgroup. For example, ADHD seems to occur more frequently in boys than girls: let us suppose in 15% of boys and in 5% of girls. If a questionnaire for detecting ADHD uses norm groups for each gender and the limit is set at the 90th percentile in each group, this means that an equal number of boys and girls are classified as having ADHD. This is an underestimate of the boys and an overestimate of the girls. In such a case, one norm table for boys and girls combined is the better choice.

The point is not to cover all possible subgroups, but to choose subgroups that are relevant to the nature and purpose of the test. These include groups based on gender, age or ethnic background.

Notes for question 4.7: 'Is there information about the accuracy of the measurement, and the corresponding confidence intervals?'
Interpretation of test scores demands information on the accuracy of the measurement and the corresponding confidence intervals. Measures providing information on the accuracy of the measurement are the standard error of measurement, the standard error of estimate and (for tests constructed according to an item-response model) the test information function/ standard error (Drenth & Sijtsma, 1990, 2006, pp. 226-235; Lord & Novick, 1968).

The standard error of measurement is calculated from the standard deviation and the reliability and is equal to $S_E = S_x\sqrt{1-r_{xx}}$. It can be used to estimate a confidence interval for the true score T. The true score T is non-observable and is easily estimated by using the observed score X (therefore, $\hat{T}=X$). This assumes that measurement errors are normally distributed. The confidence interval for T is symmetrical around the observed score X, which functions as an estimate of T. A confidence interval of 95% is obtained by calculating for a given observed score X: $X \pm 1.96S_E$. The lower limit of the interval is thus obtained by subtracting $1.96S_E$ from X, and the upper limit is obtained by adding $1.96S_E$ to X. This interval provides an impression of the accuracy of the measurement and can be used to test hypotheses about a person's true score.

The standard error of estimate can be used for the same purposes as the standard error of measurement: in other words, the estimate of a confidence interval for the true score. The difference is that another estimate of the true score is used. Here, the true score is estimated by means of linear regression, which results in the formula $\hat{T} = r_{xx} X + (1-r_{xx})\bar{X}$. In this formula, $\bar{X}$ is the mean of the group in which the reliability is determined. This formula is known as *Kelley's formula*. Because it gives a regression estimate of the true score, the standard error of estimate, $S_{est} = S \times \sqrt{r_{xx}}\sqrt{1-r_{xx}}$, is now used to estimate a (95%) confidence interval for T. This is done by calculating $\hat{T} \pm 1.96S^{est}$, while using the estimated true score from Kelley's formula. The advantage of Kelley's method over the first formula is that more information is used to estimate the true score. The idea behind this is that if the reliability is lower, observed scores receive less weight while the group mean receives more. Conversely, if the reliability is higher the group mean plays hardly any role and estimation of T is determined almost entirely by the observed score X. In this way, the reliability and the mean test score play a role in the estimate of T. This is the essence of the difference with using the standard error of measurement. The consequence of using more information is that $S_{est} < S_E$. In short, the second method is more precise than the first.

The second formula for estimating the true score (*Kelley's formula*) is often used in practice while the resulting standard error is erroneously reported as the 'standard error of measurement'. One must beware of this.

The confidence interval is important when carrying out statistical tests. We will name two possibilities: is the score of person v different to that of another person w or to a cut-off score X_o?

The first possibility encompasses the null hypothesis that the true scores of two persons, say person v and person w, are equal; thus, H_0: $T_v = T_w$. We then test whether the difference is equal to 0, and evaluate the standard error of measurement of the difference $D_{vw} = X_v - X_w$. This standard error of measurement is equal to $S_{E(D)} = \sqrt{2S_E}$. A 95% confidence interval for the true difference which is estimated by using the observed difference is equal to $D_{VW} \pm 1.96 \sqrt{S_{(ED)}}$. If the value of 0 lies within this interval, the null hypothesis is accepted, but if the value 0 lies outside the interval, the null hypothesis is rejected. For the second possibility, comparison of the test score with a cut-off value, this cut-off value must be perfectly reliable. We then only need to check whether the cut-off value is contained within the confidence interval $X \pm 1.96S_E$. When working with the standard error of estimate, one must always replace S_E in the formula with S_{est}.

Standard errors of measurement and their accompanying confidence intervals can be estimated for other scores as well, such as true difference scores. One must bear in mind that difference scores as used in contexts such as score profiles are notoriously unreliable, and the confidence intervals are extremely long. Literature gives a great deal of information on psychometric problems with difference scores, but this remains a thorny issue (see Allen & Yen, 1979, pp. 208-211; Drenth & Sijtsma, 2006, pp. 241-243; Murphy & Davidshofer, 1998, pp. 138-139).

If the test or questionnaire is constructed by means of an item-response model, the test author can choose to state either the test information function or its opposite, the standard error, depending on the scale. Using this standard error again makes it possible to estimate confidence intervals for the level of the respondent (in the context of item-response theory, this is a different parameter than the true score). The difference with confidence intervals on the basis of the classical standard error of measurement is that the confidence intervals now vary over the scale values. This makes it evident that not everyone can be measured with the same precision by the same test. It is advisable to present this local information on measurement precision in both graph and numerical table form: the former gives clarity, while the latter states the exact values.

For a positive rating on this item, the test author must use at least one of the three above-named reporting methods, standard error of measurement, standard error of estimate or test information function/conditional standard error, including a satisfactory explanation for the test user on how to use confidence intervals. It is recommended that the manual should include these intervals for every raw score or standard score.

CONTENT-REFERENCED OR CRITERION-REFERENCED INTERPRETATION

When a test uses cut-off scores, these scores are used to divide the entire score range into two or more categories. These categories may be used for descriptive purposes only, but are generally intended to make a distinction between groups of tested persons who are being offered a different programme or treatment, or for whom there are different expectations. Cut-off scores can then be used in several ways: an employer can screen for potential employees, a school can offer groups of students varying study programmes, mental health care institutions can make decisions about the indicated therapy or the presence of a particular psychopathology, and minimum passing scores can be determined for certification by the institution in charge.

There are several methods for determining cut-off scores, which are also referred to in literature as standard determination procedures. In general, we can distinguish between procedures that rely on the judgement of experts (questions 4.8 through 4.10) and procedures that rely on actual data on the relationship of the test score with one criterion or another (questions 4.11 through 4.13). In certain cases, the cut-off score can also be determined by directly referring to a percentage in the reference group. For example, a person who is among the lowest 10% of scores on a school test is eligible for extra educational facilities. Or if a person scores among the top 20% on a questionnaire for psychopathology, that person is placed in a treatment programme. Since cut-off scores like these are in fact based on comparison with a norm group, the requirements formulated in question 4.3 apply to them as well.

Content-referenced interpretation

Literature (Berk, 1986; Cascio & Aguinis, 2005; Cizek, 1996; Hambleton, Jaeger & Plake, 2000; Livingston & Zieky, 1982; Vos & Knuver, 2000) mentions various standard determination methods in which standards or norms are determined using the input of a number of raters (content experts). These methods include examinee-centred and test-centred methods. In the first category of methods, raters are asked to indicate what behaviour they could expect, from a student (real or imagined) who lies on the boundary between failing and passing (Van Berkel, 1999), for each item on a representative test. In the second group of methods, however, real persons are assessed, and the cut-off score is derived from an actual score distribution. By means of a chosen standard determination method, a norm can then be established.

Notes for question 4.8: 'Is there sufficient agreement between the raters?'

Once obtained, the norm can only be legitimised on the basis of broad agreement among the raters. When rating this question, one must establish whether the interrater agreement is being reported, and not the interrater reliability. Interrater agreement refers to identical judgements by different raters, while interrater reliability refers to relatively identical judgements by different raters, in which the absolute level of the ratings need not be equal.

Different coefficients can be used, depending on the measurement level of the data. With data at a nominal level, coefficient kappa, *K*, is often used, and with data at ordinal level, weighted coefficient kappa, K_w, is often used. For interpretation of the level of coefficient kappa, no definitive norms are mentioned in literature, although a *K* of .60 is generally deemed a minimum value for acceptable interrater agreement. Shrout (1998) refers to kappas in the range from.61 - .80 as moderate and in the range from .81 - 1.00 as substantial. The interpretation should be approached with some caution, as the prevalence or the base rate can influence the value of kappa.

To measure data at interval level, the intraclass correlation coefficient (ICC) is frequently used. The ICC is a ratio of variance components, and in the case of the interrater agreement coefficient, the error variation is formed by the variation within paired observations. When reporting this coefficient, one must assume the presence of only one rater. Shrout (1998) points out the comparability of kappa with ICC and posits that, just as with kappa, an ICC of > .80 can be interpreted as 'good'.

On this basis, the following norms can be established for the interpretation of the coefficient kappa and the intraclass correlation coefficient:

Coefficient kappa or intraclass correlation coefficient	*K* or ICC ≥ .80 .60 ≤ *K* or ICC < .80 *K* or ICC <.60	good sufficient insufficient

Notes for question 4.9: 'Are the procedures for determining the cut-off scores correct?'

To judge whether norms can be legitimised, it is important for the test author to describe precisely what procedures were followed. When rating this question, one must check whether the following conditions have been met:

- Are all steps and decisions taken in agreement with the definitions and procedures stated in the method?
- Have all the steps defined in the model been kept constant? This includes matters such as instructions, materials and the provision of statistical information on performance distributions.

If one of these two conditions is not met, the rating must be 'insufficient'. For a rating of 'sufficient', the two conditions above must be met. For a rating of 'good', the test author must additionally justify the use of a particular standard determination procedure and indicate how any inconsistencies in ratings were dealt with during the standard determination procedure.

Notes for question 4.10: 'Have the raters been properly selected and trained?'

Because the raters play a prominent role in the standard determination method, one must choose them with care. Potential raters must at least have knowledge of the domain the test is concerned with, and the rater should have preferably been trained in evaluating the work of test takers. It is also important that every rater should understand the standard determination method that is to be followed, thus avoiding any disagreements arising because raters are applying the method in different ways. This goal can also be achieved by offering a training programme. To judge whether the test author has chosen the raters prudently, a description of the selection procedure and the training offered to raters must be provided.

Criterion-referenced interpretation

Cut-off scores can be empirically supported in all sorts of ways. However, one shared feature is that in all these cases, it is not only the test takers' scores that are available, but also data on the criterion to be predicted, and thus on the test-criterion relationship. In essence, this involves research on the criterion validity, which also has the function of empirically establishing a norm. The research evaluated here concerns the latter function. A few examples:

- After research that establishes the relationship between test scores and job performance, personnel selection departments can determine minimum required scores and/or construct expectancy tables.
- In clinical psychology, ROC curves and sensitivity and specificity values based on the relationship between test scores and independently established criteria can be used to determine the most favourable cut-off scores.
- When awarding licences or diplomas, the pass/fail limit can be determined by investigating which test score shows a favourable relationship between participants who were successful in practice and those who were not.

Notes for question 4.11: 'Do the research results justify the use of cut-off scores?'

When cut-off scores are empirically supported, the researcher will have to provide proof for the usability of the chosen cut-off score. For personnel selection, this might call for data on the success ratio, while a clinical situation might require data on the sensitivity and specificity (see the notes for key questions 7.1 and 7.2). No general recommendations for the desired value of these measures can be given, not only because 'what is high or low' can vary for each criterion to be predicted, but also because the prediction results are influenced by other variables like the base rate or the prevalence. It should thus be left to the reviewer's expertise to weigh the various factors against one another and make a judgement on the values of the results obtained.

Notes for question 4.12: 'Is the research sample appropriate for the intended use?'

The research that determines the cut-off score must be related to the population for which the test is used. If the composition of the sample is more heterogeneous than that of the population for which the test is to be used, and in which decisions will ultimately be made, this will not only lead to overly biased results, but may possibly produce cut-off scores different to those that would have been obtained from research on an appropriate group. For an accurate evaluation, the sample must be described with the use of any relevant psychological or demographic variables.

Notes for question 4.13: 'Is the research sample large enough?'

Cut-off scores can be viewed as 'normal' points in a score distribution to which a special meaning is assigned. For the precision with which these points have been established, the same requirements apply as those for norm tables, whose precision is primarily determined by the group's size. Bear in mind that determining one or more cut-off scores involves only a limited number of points, while with norm tables, the precision of the entire score distribution can be affected. The requirements set for the size of the research sample can therefore be relaxed in comparison with the requirements in force for norm-referenced interpretation (see notes for question 4.3.a). Assuming that cut-off scores are only determined in situations having to do with 'important decisions on an individual level' (see the recommendations for question 4.3.a for a description), a research sample consisting of at least 300 persons is rated 'good', a research sample of at least 200 persons is rated 'sufficient' and a research sample of less than 200 persons is rated 'insufficient'.

Rules for determining final rating for criterion 4 **Norms** **Norm-referenced interpretation**		
All three key questions (4.1, 4.2 and 4.3) are rated '3'.	Sum score 4.4 through 4.7* ≥ 9	good
	Sum score 4.4 through 4.7* = 7 or 8	sufficient
	Sum score 4.4 through 4.7* ≤ 6	insufficient
Key question 4.1 is rated '3' and both key questions 4.2 and 4.3 are rated at least '2'.	Sum score 4.4 through 4.7* ≥ 9**	sufficient
	Sum score 4.4 through 4.7* ≤ 8***	insufficient
One or more of the three key questions is rated '1'.		insufficient
* For questions 4.7 a, b, and c, the highest-scoring sub-question is taken. ** If questions 4.4 through 4.7 are all rated '3', the final rating is 'good'. *** If questions 4.4 through 4.7 are all rated '2', the final rating is 'sufficient'.		
Content-referenced interpretation		
Key questions 4.1, 4.2 and 4.8 are all rated '3'.	Sum score 4.9 and 4.10 ≥ 5	good
	Sum score 4.9 and 4.10 = 3 or 4	sufficient
	Sum score 4.9 and 4.10 = 2	insufficient
Key question 4.1 is rated '3' and key questions 4.2 and/or 4.8 are rated at least '2'.	Sum score 4.9 and 4.10 ≥ 5	sufficient
	Sum score 4.9 and 4.10 ≤ 4	insufficient
One or more of the three key questions is rated '1'.		insufficient
Criterion-referenced interpretation		
Key questions 4.1, 4.2 and 4.11 are all rated '3'.	Sum score 4.12 and 4.13 ≥ 5	good
	Sum score 4.12 and 4.13 = 3 or 4	sufficient
	Sum score 4.12 and 4.13 = 2	insufficient
Key question 4.1 is rated '3' and key questions 4.2 and/or 4.11 are rated at least '2'.	Sum score 4.12 and 4.13 ≥ 5	sufficient
	Sum score 4.12 and 4.13 ≤ 4	insufficient
At least one of the three key questions is rated '1'.		insufficient

5 Reliability

Classical test theory posits that a test score (X) is constructed from a reliable portion, also called the true score (T), and a portion that is attributable to the influence of random measurement errors. This latter portion is called the measurement error (E). The test score is the sum of the true score and the measurement error: X = T + E. It would be ideal to measure only the true score, but the fact is that test scores do also contain measurement errors. The goal of reliability analysis is to estimate the influence of measurement errors on the test scores.

The variance of the test scores in a group of persons (S_x^2) is constructed of true variance (S_T^2) and error variance (S_E^2), so that $S_x^2 = S_T^2 + S_E^2$. In its most basic form, error variance represents the distribution that results from random measurement errors, so that the true variance represents all systematic differences between respondents. Parallel forms reliability is the ratio of this reliable variance and the variance of the test scores.

Besides the interpretation of measurement errors as random score components, there is another interpretation that states that measurement errors contain all unintended components of the test score, beginning with the random ones, but subsequently the unintended systematic ones as well. In this case, the true variance represents the distribution of the intended score components, and the error variance represents the distribution that results from the unintended components, including the random measurement errors. An estimate of this is obtained by using techniques from generalisability theory, item-response theory and structural equation models. In both cases, it is important, but not always strictly necessary, to analyse data specially collected for this purpose.

An example of a test that measures not just the intended property but another property as well is a calculation test whose variance in the test scores is dependent not only on calculation skills (intended) but also language skills and chance (both unintended). The first form of reliability is equal to the ratio of the variance as the result of differences between respondents in calculation skills and language skills together, and the variance in the test scores. The second form is equal to the ratio of the variance of only the intended calculation skills, and the variance in the test score.

The sources of error variance can vary greatly and do not have to be related only to unintended psychological properties like the language skills in the previous example. An alternative possibility is that one investigates the extent to which a test score can be repeated over a specific period. In this way, 'mood' can be rated at the same time point with two questionnaires functioning as parallel instruments, and it can prove that the measurement at that time point was very reliable. However, if there is a long time interval between two administrations of the same test (not two different versions), the correlation between the two test scores may prove to be low. One can then conclude that the differences between respondents measured over a long period were only for a small portion systematic. The reliability, or the test-retest reliability in this case, is therefore too small for generalisation of the test score over the investigated time period.

The indices for reliability that mention the error source thus make it possible to declare a test reliable for a particular purpose. The use of the traditional reliability indices as explained in question 5.2 makes it possible to establish the generalisability of scores over versions (parallel forms reliability; reliability estimates on the basis of inter-item relationships give an estimate of this, which we will discuss later), time points (test-retest reliability) and raters (inter-rater reliability). This summary confirms a fact to which we have already alluded: it is not possible to speak of the definitive reliability of a test. Rather, different forms of reliability are distinguished depending on the nature of the variance analysed in a particular reliability study.

It is also important to acknowledge that the outcomes of the reliability study for a particular test depend on the group being studied. If the same characteristic is measured in two groups, the reliability is greatest in the group with the widest variance in the test scores. On the other hand, if the test measures only the intended characteristic in one group, and both the intended characteristic and an unintended property in the other group (think of the earlier example of calculation and language skills), then the validity of the test is at risk, and it is not advisable to compare the scores of people from different groups with one another.

Although a test frequently contains more than one component (scales, subtests), the reviewer usually will give one rating for the reliability, which supplies a summary of results for the various components. This is the case with questionnaires consisting of various scales such as the *BIT*, the *EPPS* and the *NPV*, and for test series that consist of several subtests that can in principle be administered separately, such as the *DAT*, the *DVMH* and the *MCT-M*. In such cases, the lowest coefficient is decisive for the rating. However, when this coefficient is a clear negative exception (for example, every subtest but one is 'good' and one is 'insufficient'), the higher rating may be given (in this case, 'good').

This exception should be mentioned in a footnote. Another situation can arise when the scores on the subtests are added up to arrive at a total score, as is done with some intelligence tests. Here, there are three possible approaches:

- If only the interpretation of the total score matters, of course only the reliability of this score needs to be rated.
- If the test author states that the total score is indeed the most important, but that the interpretation of subtest scores is also possible, the reliabilities of the subtest scores should be rated with the criteria that apply one level below the level that applies to the total score (see notes for question 5.2). If the total score is categorised as 'important', subtest scores should be in the category 'less important'. In most cases, subtest scores are less reliable than total scores, but when the above rule is applied the ratings can be equal.
- If the test author makes no distinction between the importance of the total score and the factor or subtest scores, they are rated in the same way, as equally important.

When the ratings given for the reliability of factors/subtests and total scores differ, this should be mentioned in a footnote to the rating. It is also important to note that only one rating is given when a test author computes the reliability of several groups, and these are found to differ. Here, the result for groups that represent the primary intended use should be weighted more heavily. Mutatis mutandis, such cases are also governed by the rule stated above. The lowest reliability coefficient is decisive for the rating except in cases where this value is a clear exception.

Questions for criterion 5 Reliability					
		ins.	suf.	good	
Key question 5.1	Is there information about the reliability of the test? If the rating of this question is negative, proceed directly to criterion 6.	1		3	
5.2	Are the outcomes of the reliability research sufficient with respect to the type of decisions that are based on the test?				
5.2.a	Parallel forms reliability	1	2	3	Not applicable
5.2.b	Reliability on the basis of inter-item relationships	1	2	3	Not applicable
5.2.c	Test-retest reliability	1	2	3	Not applicable
5.2.d	Inter-rater reliability	1	2	3	Not applicable
5.2.e	Methods based on item-response theory	1	2	3	Not applicable
5.2.f	Methods based on generalisability theory or structural equation models	1	2	3	Not applicable
5.3	What is the quality of the reliability research?				
5.3.a	Are the procedures for calculating the reliability coefficients correct?	1	2	3	
5.3.b	Are the samples for calculating the reliability coefficients consistent with the intended use of the test?	1	2	3	
5.3.c	Do the data provided make it possible to make a well-grounded judgement on the reliability of the test?	1	3	3	

Notes for key question 5.1: 'Is there information about the reliability of the test?'

This refers to reliability coefficients and the results of generalisability research. One can also use item-response theory to report a reliability coefficient, a table or figure with standard errors or an information function.

Notes for question 5.2: 'Are the outcomes of the reliability research sufficient with respect to the type of decisions that are based on the test?'

No general statement about the desired size of a reliability coefficient or a similar measure can be made, because the purpose for which the test is used must always be taken into account. Nunnally and Bernstein (1994, p. 265) indicate that a test that is used for important decisions must have a reliability of at least .90. Important decisions are understood as decisions taken on the basis of the test scores that are essentially, or in short term, irreversible, and are for the most part beyond the influence of the test taker. Taking this value as a basis, the following rules have been formulated in the table below.

Tests for important decisions at individual level (for instance, personnel selection, placement in special educational programmes, admission to/discharge from clinical treatment).
good: r > .90
insufficient: r < .80
sufficient: .80 < r < .90
Tests for relatively less important decisions at individual level (for instance, evaluating progress and general descriptive use of test scores in such areas as vocational guidance and admission to therapy).
good: r > .80
insufficient: r < .70
sufficient: .70 < r < .80
Tests for research at group level (for instance, measurement of team satisfaction, atmosphere in the classroom, or organisational culture).
good: r > .70
insufficient: r < .60
sufficient: .60 < r < .70

When estimating variance components in generalisability studies, thresholds must be observed that are equivalent to the values above. This is also true for a reliability coefficient based on item-response theory and structural equation models. For standard errors and the information function, it is more difficult to devise simple rules of thumb. This subject is further discussed under 5.2.e.

Very often, more than one of the indices under 5.2.a through 5.2.f will be mentioned for one test. When deciding on the final rating for reliability, the coefficient that must be weighted most heavily is the one that best matches the purpose for which the test is being used. For prediction over time, for example, a stability index is the first priority.

Notes for question 5.2.a: 'Parallel forms reliability'

Parallel forms reliability can be used to estimate the reliability as the ratio of systematic variance and variance of the test scores. Tests are parallel when their scores for the same group show the same means, variances, and correlations with other variables. If these features are present, the correlation between the test scores is then equal to the reliability of the individual tests. If the test versions are not parallel, their correlation gives an underestimate of the parallel forms reliability. This correlation can then also be seen as a measure of the generalisability over differing, non-parallel test versions.

Parallel forms reliability can be useful for pure speed tests. The correlation between the test halves, formed on the basis of either halving the testing time or the test content, can be considered as parallel forms reliability, but then only for half a test. A correction for test length can subsequently be applied to obtain an estimate of the reliability of the entire test.

Notes for question 5.2.b: 'Reliability on the basis of inter-item relationships'

Cronbach's (1951) alpha is based on the covariance between the items on a test and is almost always used to estimate the reliability of the test score. Three factors are important here. First of all, alpha is a minimum value for parallel forms reliability. The value of alpha is therefore lower than the true reliability of the test (Novick & Lewis, 1967). Second, there are many alternative methods that strongly resemble alpha and sometimes give an estimate of the reliability that is closer to the true reliability than alpha's. One such example is Guttman's lambda2 (Guttman, 1945). The value of Guttman's lambda2 is always closer to the reliability than that of alpha, although the differences are small. Third, a great deal of literature on test theory reports that alpha is a measure for the internal consistency of a test. This is the origin of the often-used designation 'internal consistency coefficient'. This means that a higher value for alpha would indicate that the items possess the same property to a higher degree. However, psychometric literature has shown that this interpretation is incorrect: a low value of alpha can be associated with both a one-factor and a multiple-factor test, and the same can be said of a high value of alpha (Drenth & Sijtsma, 2006). If one wants to make a statement about the test composition, the best alternatives are techniques like factor analysis and principal component analysis.

In addition, the estimate of the reliability is sometimes calculated with the split-half coefficient or split-half reliability. This method is discouraged, as the outcomes are dependent on the random distribution of items over test halves. Besides, the mean of the split-half coefficients based on all possible divisions of the test into two halves is identical to Cronbach's alpha (Lord & Novick, 1968). For this reason, it is probably better to use alpha or lambda2.

Another method that is rarely used is the greatest lower bound (GLB; Ten Berge & Sočan, 2004). Under the assumptions of classical test theory and given a table with inter-item covariances, this method searches for the lowest possible value for reliability. The result is a worst-case scenario for reliability that always appears to be greater than Cronbach's alpha and results of comparable methods. In view of the fact that the GLB is closer to the true reliability, it is a better underestimate of it than the other methods described above.

For speed tests and tests with so-called heterogeneous scales, measures for internal consistency such as Cronbach's alpha are not useful. The same applies to causal indicators (see Nunnally & Bernstein, 1994, for the difference between causal and effect indicators) and to the comparable so-called emergent traits (for the distinction between emergent traits on the one hand, and single or multi-faceted traits on the other, see Schneider & Hough, 1995), for which it is similarly unnecessary for the items to be correlated to each other. In all these cases, one of the other reliability indices may provide the answer. For pure speed tests, the parallel forms method (see also the comment on question 5.2.a) or the test-retest method may be used. However, many power tests also have a time limit. Particularly when a significant percentage of the test takers have not been able to complete the last items of the test, internal consistency should not be calculated automatically, because this can cause reliability to be overestimated. In such cases, an estimate of reliability can be obtained by splitting the test into two halves of equal test length (for instance the odd and the even items) and correcting the correlation between the scores on these halves (which have been administered within half the testing time) for test length. When speed is not a primary factor, say, if no more than 30% of the test takers fail to complete the last item, another possibility is to apply a correction formula for reliability (De Zeeuw, 1978). Another possibility in that case would be to estimate the reliability for only those items completed by at least 90% of the test takers.

For heterogeneous scales and causal indicators, the test-retest method can be used, but for these kinds of tests correlations with other variables may replace reliability indices. For causal indicators in particular, a thorough specification of the domain is essential (see Chapter 1, Theoretical basis of the test construction).

For tests that use starting or breaking-off rules, and for adaptive tests in general, Cronbach's alpha cannot be used indiscriminately. Some authors wrongly assert that this is possible because simulations show a high correlation between the adaptive score and the score for the whole test. In this case IRT models have to be applied, or a method such as that employed by Laros & Tellegen (1991). In this method, reliability is estimated on the basis of various breaking-off rules and the correlations of these scores with a criterion variable.

Notes for question 5.2.c: 'Test-retest reliability'

Generalisability over time is estimated by means of test-retest correlations. Here, a test is repeated in the same research sample. Both the time interval and any relevant events which took place during the interval must be reported with maximum accuracy. The length of the time interval and the value of the correlation determine the degree to which the test performance can be generalised over time. Test-retest coefficients are especially desirable when the test is intended for prediction over time, but also when one expects that the construct to be measured is related to age, as in intelligence tests for children.

Notes for question 5.2.d: 'Inter-rater reliability'

For observation and rating scales in particular, it is important to know whether scores can be generalised with respect to observers or raters. Measures that can be used are agreement indices such as Cohen's kappa (Cohen, 1960,1969), Gower's coefficient (Gower, 1971), the identity coefficient (Zegers & Ten Berge, 1985) and other measures that account for differences in means and variances between raters (see Zegers, 1989 for an overview). Study of the variance components or the factor structure of the observers' or raters' behaviour may also be relevant here.

When rating the reported values, it is important to consider the kind of coefficient used. For example, literature distinguishes between inter-rater *agreement* and inter-rater *reliability* (Heuvelmans & Sanders, 1993). The difference is that in the denominator of the formula for the inter-rater reliability, the variance component for the raters is left out. This coefficient will therefore give higher outcomes than the formula for inter-rater agreement. The differences in the transformations applied to the scores for the various coefficients mentioned by Zegers (1989) are similar. Finally, it must be stressed that high inter-rater reliability is an essential condition for high test reliability, but that it is not the same thing as high test reliability.

Notes for question 5.2.e: 'Methods based on item-response theory'

Literature on item-response theory mentions two approaches for establishing the accuracy of a test score. The first approach is closely aligned to the classical definition. There are two methods. The first method states the reliability of the estimated latent trait, which in item-response theory replaces the estimated true score, in other words, the test score (Embretson & Reise, 2000). The second method is known by the name *rho* and was proposed by Mokken (1971). This method is based on information on individual items and yields an estimate of the reliability of the test score when certain conditions are met that are typical of item-response models. The rules for interpretation of both reliability methods are the same as the rules mentioned in the notes for question 5.2.

The second approach is fundamentally different to what we are accustomed to from classical test theory, because it gives an estimate of the accuracy as a function of the scale of the latent trait. The result is a function instead of a coefficient. This function is the so-called test information function, of which variants can be stated. The test information function can also be converted into a function that gives the standard errors that belong to the estimate of the latent trait values. In this way, every latent trait value can be used to calculate confidence intervals for the true latent trait value. When there are different latent trait values, there are also confidence intervals of different lengths. This shows that not every scale value, and therefore not every person, is measured with equal accuracy. For example, if most of the items for the group to which the test is administered are of average difficulty, the confidence intervals in the middle of the scale, where the precision of measurement is relatively high, are shorter than at the extremes. This illustrates that it is not possible to precisely measure people for whom all items are either difficult (and had a low test score) or easy (high test score). The items are simply not suitable for these individuals. In order to determine such confidence intervals, classical test theory uses the standard error of measurement and assumes that this applies to every individual. This implies that the confidence intervals are equally long for everyone, regardless of their position on the scale.

Concrete recommendations for the level of information functions or the length of confidence intervals are difficult to give, because they depend on the application of the test and the importance of the decisions that must be made on the basis of the test scores. One would do well to consult the literature on item-response theory. For instance, see Reise and Havilund (2005, p. 234) who present a way of using the information function, and Langenbucher et al. (2004), who clarify that a scale cannot measure with equal reliability in every location.

Notes for question 5.2.f: 'Methods based on generalisability theory or structural equation models'

Finally, we mention the possibility of estimating reliability by using structural equation models. A major role is played here by confirmatory factor models. At the moment, this method is rarely employed, but see Raykov (1997) and Green and Yang (2009).

Notes for question 5.3.a: 'Are the procedures for calculating the reliability coefficients correct?'

Some points of special attention for each of the forms of reliability we have mentioned are stated below:

- When parallelism of two tests cannot be made plausible (the most critical feature here is identical correlation behaviour with other variables), the calculated coefficients should instead be considered as an index of convergent validity.
- When constructing a test or scale, there is often an attempt to obtain the highest Cronbach's alpha possible. This is often done by using items with homogeneous content. This can lead to very specific test content, in which the construct being measured is far narrower than the one originally intended. This does not always produce a useful test or scale. The fact that a subgroup of items in a test shows higher intercorrelations than the rest of the items, or even the existence of several such subgroups within a scale, need not preclude a high Cronbach's alpha or a coefficient related to it. On the contrary, if correlations with the other items are moderate, such homogeneous subgroups enhance the value of these estimates of reliability. Relatively high inter-item correlations within a subgroup of items can arise because these items share unintended variance not common to the other items in the test, for instance when items are formulated similarly or have a specific word in common. Such unintended variance contributes to higher estimates of reliability because the variance of the systematic differences between respondents is greater. Unintended narrowing of the construct can be avoided as early as the developmental phase of the test by testing for unidimensionality, for instance by using structural equation models (programmes such as *AMOS*, *Mplus* or *LISREL*), and using the theoretical basis of the test construction to take appropriate action. This last remark is important because the test author can explicitly state whether he wishes to measure a narrow or a multi-dimensional construct. The usefulness of such a starting point and the outcomes of the research for the dimensionality of the test are evaluated elsewhere in this document. At this stage, the COTAN reviewer is only asked to assess the reasons for a high Cronbach's alpha in the light of this discussion, and to be especially aware of the effects mentioned above if analyses for unidimensionality have not yet been performed.
- No strict guidelines can be given for the length of the test-retest interval. As a rule, a very short interval (up to a few weeks) is not suitable because of the effect of memory. A very long interval (longer than a year) may not be useful either, because external events or experiences may exert strong influence on the person and on the retest score as well. In such a case, the instrument can in fact no longer be considered reliable. However, the intervals mentioned above are rather arbitrary and must be viewed in relation to the age of the group tested and the nature of the test. Additionally, the purpose of the test plays a part in the choice of the test-retest interval. For example, with a test that is intended for long-term prediction, it is wise to choose a relatively long test-retest interval.
- When inter-rater reliability is used to estimate the reliability of only one rater's judgement, observations or ratings must be carried out independently. This fact should be clear from the description of the research design.

Notes for question 5.3.b: 'Are the samples for calculating the reliability coefficients consistent with the intended use of the test?'

Reliability coefficients must be computed for the groups for which the test is used. This implies that they have to be computed per norm group, since the scores of test takers are compared with such a group and it is the reliability of the measurement within this reference group which matters. It is therefore incorrect, and even misleading, to calculate reliability coefficients for the total of the groups or, as sometimes happens, for a selection of groups at extremes of the distribution. Since the size of the reliability coefficient also depends on the distribution of the scores, the coefficient calculated for the scores of the total group will almost always be higher, certainly for the extremes, than for each norm group separately. If the coefficients for each norm group have not been calculated, a rating of 'insufficient' must be given.

Notes for question 5.3.c: 'Do the data that are provided make it possible to make a well-grounded judgement on the reliability of the test?'

Below are some examples of information which must be available in order to evaluate the quality of the reliability study:

- Are the standard deviations of the scores of the test and the retest group given?
- For tests with a time limit, is it stated for each item what percentage of test takers have answered that item?
- Have the samples for which the reliability coefficients are calculated been described in sufficient detail?
- Is it mentioned how many observers or raters are included in the reliability coefficient being reported?
- Observers or raters usually receive training for their job. This training will influence the quality of the ratings and therefore the level of inter-rater reliability. The description of the training programme should be detailed enough to enable new test users to prepare themselves in the same way so that the reliability of the ratings can be generalised to other situations. It must be feasible for new users to acquire the same skill level. It is also important to mention whether the reported reliability coefficient relates to the assessment of a single observer or rater or the averaged assessment of several observers or raters.

In an extreme case where no descriptive information at all on the reported reliability coefficients is provided, this question can be rated 'insufficient'. In most cases, enough information will be provided to allow the quality of the reliability research to be rated. Especially in borderline cases (insufficient/sufficient, sufficient/good), inadequate information can be a reason for giving the lower rating.

Rules for determining final rating for criterion 5 Reliability			
The key question is rated '3'.	Question 5.2 is rated '3'.	Question 5.3 is rated '3'.	good
		Question 5.3 is rated '2'.	sufficient
		Question 5.3 is rated '1'.	insufficient
	Question 5.2 is rated '2'.	Question 5.3 is rated '3'.	sufficient
		Question 5.3 is rated '2'.	sufficient
		Question 5.3 is rated '1'.	insufficient
	Question 5.2 is rated '1'.		insufficient
The key question is rated '1'.			insufficient
If there is a positive rating for question 5.1, question 5.2 on the size of the reliability coefficient then produces a provisional rating. This provisional rating may be lowered on the basis of the answer to question 5.3 on the quality of the research performed			

6 Construct validity

Validity is the extent to which a test fulfils its purpose. Can the intended conclusions be drawn from the test scores? Literature mentions many forms of validity: Drenth & Sijtsma (2006, pp. 334-340), for example, mention eight different forms. These distinctions reflect the purpose of the validity research or the validation process using specific data-analysis techniques. Recent decades have shown a strengthening of the standpoint that different forms of validation determination should not be seen as different forms of validity, but as different ways of collecting information on the validity, and that validity ought to be seen as a unitary concept (see *Standards for Educational and Psychological Testing*, 1999). From this standpoint, the most important task is to collect validity information that matches the purpose of the test, such as description, prediction or classification. Validity thus refers to scientifically grounded argumentation to support a particular interpretation of a test, where not all types of evidence are equally important to the purpose (Ter Laak & De Goede, 2003). In other words, we are not looking for a property of a test, but a property of the interpretation of test scores. A more recent viewpoint on validity comes from Borsboom, Mellenbergh and Van Heerden (2004), who posit that validity is concerned with the question of whether the attribute being measured is capable of causing variance in the outcomes of the measurement. In this approach as well, a distinction between types of validity is not the issue.

Whatever approach to validity one chooses, for a standardised review one must apply a degree of structure to the validity concept. This is in line with the classical three-category classification for the purpose of validity research as given in such publications as *Richtlijnen voor ontwikkeling en gebruik van psychologische tests en studietoetsen (Guidelines for development and use of psychological tests and educational exams)* (Evers et al., 1988): content validity, construct validity and criterion validity. Of these three, validity information related to the relevance of a test's content (content validity) and to the meaning of a test score (construct validity) is considered important for all types of tests, regardless of their purpose. However, this does not apply to information on the predictive value of test scores (criterion validity). For tests without any predictive function, such as those for evaluating educational progress, this type of information is not required. On the other hand, it is true that data on criterion validity can be employed to rate the construct validity (question 6.2), because these data can also help to clarify what is being measured by the test. In that case, data concerned with criterion validity are in fact also part of the process of construct validity (see Anastasi, 1986; Messick, 1988).

In the present review system, data or arguments about content validity are treated as a component of the test development process and have therefore already been discussed in Chapter 1 'Theoretical basis of the test construction'. The present chapter is devoted to construct validity, while criterion validity is discussed in Chapter 7.

Construct validity is intended to investigate whether the test does indeed measure the property it is meant to measure. Does the test measure the intended construct, or does it partly or mainly measure something else? Frequently used methods or techniques to provide evidence of construct validity are performing a factor analysis to demonstrate unidimensionality, comparing the mean scores of groups that are expected to differ, and calculating correlations with tests that are supposed to measure the same construct (convergent validity). This kind of research is normally quite easy to perform and the results can give an initial indication of the evidence of construct validity, but by themselves none of these indications are enough to justify a rating of 'sufficient'. Only the accumulation of such evidence, or more extended structural research or *multi-trait-multi-method* research (Campbell & Fiske, 1959), can produce a rating of 'sufficient' or 'good'.

Questions for criterion 6 Construct validity				
		ins.	suf.	good
Key question 6.1	Is there information about the construct validity of the test? If the rating of this question is negative (1), skip the other questions for this criterion and continue with criterion 7.	1		3
6.2	Do the outcomes sufficiently confirm that the intended construct is being measured (or, do the outcomes make sufficiently clear what is being measured) on the basis of information on: a. The dimensionality of the scores? b. The psychometric quality of the items? c. The invariance of the factor structure and possible item bias in different groups? d. The convergent and discriminant validity? e. Differences between relevant groups? f. The basis of other data?	 1 1 1 1 1 1	 2 2 2 2 2 2	 3 3 3 3 3 3
6.3.a	Are the procedures used to calculate the construct validity coefficients correct?	1	2	3
6.3.b	Are the samples used in the research on construct validity consistent with the groups for which the test is intended?	1	2	3
6.3.c	What is the quality of the other measures used in the construct validity research?	1	2	3
6.3.d	Is the quality of the research, as rated in questions 6.3.a through 6.3.c, good enough to confirm the rating of the construct validity as given in question 6.2?	1	2	3

Notes for key question 6.1: 'Is there information about the construct validity of the test?'
This concerns the internal or external structure of the test. The internal structure can be investigated by determining measures of association between (groups of) items or subtests, and between subtests and the test as a whole. The external structure is usually investigated by determining the relationship with other tests (convergent and discriminant validity) and calculating differences between relevant groups.

Notes for question 6.2: 'Do the outcomes sufficiently confirm that the intended construct is being measured (or, do the outcomes make sufficiently clear what is being measured)?'
Construct validity is primarily concerned with the accumulation of research evidence. Construct validation research is never completed. It may seem obvious, but still it has to be stressed here that the mere fact that construct validity is being researched does not automatically lead to a rating of '2' or '3'. For the rating, only the quality of the outcomes in the light of the theoretical basis of the test construction plays a role, of course in addition to the quality of the procedures and the research design used (see question 6.3).

For the rating of construct validity, the following six types of research data are relevant:
- Data on the dimensionality of the scores
- Data on the psychometric quality of the items
- Data on the invariance of the factor structure and possible item bias in different groups
- Data on the convergent and discriminant validity
- Data on differences between relevant groups
- Other data

Here are some notes on each of these types of research data.

Data on the dimensionality of the scores
Here, dimensionality on both test and subtest level is of interest. The research data should provide an answer to the following questions:
- When theoretical considerations call for the assumption of different sub-constructs, do these also manifest as independent factors?
- Do the scores on the test (or on subtest level, if applicable) prove to be unidimensional?
- How high is the correlation between subtests: are the constructs that are being measured distinguishable from each other?

Data on the psychometric quality of the items
The quality of the items can be rated in various ways. It is customary to consider the means of the item scores per group and at the same time to report data on the connection between items and tests or subtests. Tests based on classical test theory focus on the correlations of an item with the total score on the other items in the same (sub)test, also referred to as item-rest correlation (in *SPSS* this is known as the *corrected item-total correlation*). Tests based on item-response theory, on the other hand, focus on the fit of items within the chosen model. Depending on the model used, the following data must be supplied.

- ***Item-rest correlations***
  The size of the correlation indicates the extent to which the item in question measures the same construct as the other items, but this interpretation has its risks. The reason for this is that it must be confirmed by means such as factor analysis whether the items show high loadings on the same factor. Even if the test content is heterogeneous, an item can still have a high correlation with the total score on the other items, but because these represent a 'mixture' of various traits, the interpretation of the correlation is unclear or disputable. Another interpretation of the item-rest correlation is that of discriminatory power. Suppose that the total score on the other items creates a scale: then a high item-rest correlation means that persons with a low item score generally have a low scale score, and persons with a high item score have a high scale score. The item is thus well capable to differentiate between (groups of) people. For the rating of r_{it} values in tests featuring a high degree of internal consistency (see notes for question 5.2.b for exceptions), one can use the guidelines in the table below (based on Veldhuijzen, Goldebeld & Sanders, 1993).

r_{it} value	Rating
0.30 and higher	good
0.20 – 0.29	sufficient
0.19 and lower	insufficient

  Bear in mind that the table above is meant for r_{it} values (item-total correlations); the more customary values (item-rest correlations) can turn out somewhat lower, especially for short tests. The length of a test also seems to have an influence on the r_{it} value: the longer the test, the lower the mean r_{it} value generally is.
- ***Item parameters according to an item-response model***
  Item-response models are often used to estimate item difficulties and item discrimination values obtained on scales that are substantially different from the more familiar item means and item-rest correlations. When reporting item indices, which are typical of item-response theory, it is advisable to set them alongside the more familiar, classical item indices such as item means and item-rest correlations.

  The accuracy of the item parameter estimates in certain cases can also be rated by looking at the relationship between the standard error of the difficulty parameter $se(b_i)$ and the standard deviation of the skills distribution of the calibration population $sd(\theta)$. Here, it must hold that $se(b_i) < c*sd(\theta)$, in which c is a constant. For the rating of the standard error of $se(b_i)$, follow the guidelines in the table below.

c	Rating
$c \geq 0.5$	large (= 'insufficient')
$0.3 \leq c \leq 0.4$	medium (= 'sufficient')
$c \leq 0.2$	small (= 'good')

- ***Size of the sample***
  The sample must be sufficiently large to prevent the item parameters from being imprecisely estimated. For two reasons, there are no definitive rules that can be set up for this purpose. First, the minimum required sample size is dependent on the choice of item-response model; second, literature provides few recommendations concerning the desired sample size. It is often a question of 'experience'. Literature mentions scarcely any guidelines for the sample size needed with logistic models for dichotomous items. The guidelines in the table below are derived from research by Parshall, Davey, Spray and Kalohn (1998).

Model	N
3-parameter	N > 700
2-parameter	N > 400
1-parameter	N > 200

- ***Fit of the statistical model***
  All statistical methods are based on assumptions about distributions of variables (such as normal) and relationships between variables (such as linear). This is true of factor models and item-response models, but also for the well-known product-moment correlation. Any statements on the quality of tests based on statistical calculations can only be trusted if there is proof that these assumptions have been satisfied for the application in question. It is impractical to explain the implications of this for every technique, but the test author may certainly be expected to report the necessary information on the fit of the model in the manual of the test or questionnaire.

Data on the invariance of the factor structure and possible item bias in different groups
This research can be performed on the basis of models and procedures which are consistent with classical test theory or within item-response theory. If differences in factor structure are established or item bias is demonstrated, the consequences must be indicated, for example an estimate of the effect on the total test score. An additional benefit of the research on item bias is that it supplies information on the possible multidimensionality of the construct being measured.

Data on the convergent and discriminant validity
Both types can be obtained in one study with the multi-trait-multi-method approach. Data on the convergent validity can also be obtained via correlation with congruent tests. Data on discriminant validity are important for ruling out the accidental measurement of a construct other than the one intended. This makes it possible to measure job satisfaction without measuring negative affect, or to measure calculation skills without involving language skills to a significant extent.

Data on differences between relevant groups
Depending on the constructs intended to be measured and the features of particular groups, one may expect differences between these groups. For example, one would expect that students in grade 8 would score higher on a calculation skills test than students in grade 6. Similarly, it is expected that children diagnosed with ADHD will score higher on a hyperactivity test than 'normal' children. Group-comparing research of this type is important, because it can provide the first indication that the test is able to differentiate between groups as it is meant to. If it should unexpectedly prove that there are no differences, it is then highly unlikely that the test is measuring the intended construct. The reverse is not true, however: if there are differences between relevant groups, this is no guarantee that the test genuinely measures what it intends to measure. The calculation test can still measure language skills and the test for hyperactivity can still measure one or more forms of socially undesirable behaviour.

Other data
These might be data on the criterion validity that simultaneously provide information on the construct validity.

The question about the total score can be rated '2' if results on at least two of the research types mentioned above are reported, if these outcomes generally support the desired structure, and if they concern both the internal and external structure. A rating of '3' may be given if results on at least three of the research types mentioned above are reported, these outcomes unanimously support the desired structure, and these concern both the internal and external structure.

Notes for question 6.3.a: 'Are the procedures used to calculate the construct validity coefficients correct?'
The research design and analysis techniques used must be explained with sufficient clarity. Insufficient information can result in a rating of '2' or even '1' for this question.
As a consequence of the diversity of this kind of research, hardly any general guidelines can be given, other than the fact that the size of the research sample is important to the evaluation of the research results. A few specific points of attention are as follows:

- When the relation between items and tests or subtests is being studied, one must correct for the proportion of the item itself in the test scores, because the calculation of the value will otherwise be biased. In other words, so-called item-rest correlations must be mentioned instead of item-total correlations.
- In research on convergent validity, one should beware of the interpretation of research results for which there are no specific expectations. Research of this type can easily degenerate into 'fishing': post hoc, it is always possible to find interpretable relationships of some kind when test scores are correlated with the scores of a great number of other variables which happen to be available. In such situations, some of the significant correlations may well be a matter of chance. This chance of coincidental correlations increases in proportion to the number of subtests or scales in the test being validated.

Notes for question 6.3.b: 'Are the samples used in the research on construct validity consistent with the groups for which the test is intended?'
Research on validity must be related to the population for which the test is used. The primary issue here is the variance in the test scores of the research sample. Because validity coefficients generally are lower when variance diminishes, a validity study performed on a group that is more heterogeneous than the intended group will show biased results. It is therefore incorrect to validate a test on a cross-section of the general public if that test was intended for therapy selection among people who voluntarily requested it. To judge this, the research sample must be described with the use of any relevant psychological or demographic variables.

If the manual states that the test is intended for use in various situations and/or for a variety of groups, research must be performed in a number of these situations and/or in multiple samples.

Notes for question 6.3.c: 'What is the quality of the other measures used in the construct validity research?'
The reliabilities of the measures used must be known. It will be obvious that validating the test score using measures with a low reliability (lower than .60) is not useful, because the results will be ambiguous. Moreover, validating a test with a congruent test is only useful if the validity of the other test has itself been sufficiently investigated.

Notes for question 6.3.d: 'Is the quality of the research, as rated in questions 6.3.a through 6.3.c, good enough to confirm the rating of the construct validity as given in question 6.2?'
A negative rating ('1') for one of the questions 6.3.a through 6.3.c results in a rating of '1' for question 6.3.d. This means that the rating for the results of the construct validity research as given in 6.2 must be adjusted downwards. Multiple ratings of '2' on the questions 6.3.a through 6.3.c can also mean that the research is so flawed that question 6.3.d receives a negative rating, and that on this basis the rating for question 6.2 must be adjusted downwards.

Rules for determining final rating for criterion 6 Construct validity			
The key question is rated '3'.	Question 6.2 is rated '3'.	Question 6.3.d is rated '3'.	good
		Question 6.3.d is rated '2'.	sufficient
		Question 6.3.d is rated '1'.	insufficient
	Question 6.2 is rated '2'.	Question 6.3.d is rated '3'.	sufficient
		Question 6.3.d is rated '2'.	sufficient
		Question 6.3.d is rated '1'.	insufficient
	Question 6.2 is rated '1'.		insufficient
The key question is rated '1'.			insufficient

7 Criterion validity

Criterion validity investigates the extent to which the test score is a good predictor of non-test behaviour (retrospective, concurrent or predictive). It is important that the goal of a test as formulated (see Chapter 1) serves as a basis for specifying expectations of the type of criteria with which relationships are assumed. This is especially important when a test is comprised of various subtests or subscales; see the remarks under question 6.3.a about 'fishing'. For that matter, there is no need to demonstrate the validity of all subtests or scales to give a rating of 'sufficient' or 'good', because one single highly valid scale can already make the test a valuable instrument.

In principle, research on criterion validity is required for all types of tests, because the ultimate goal of tests is to make predictions. However, if a test manual explicitly states that the test has no predictive ambitions, and this is plausible, as with tests of educational progress, the criterion validity may then be declared 'not applicable'. In such cases, the following footnote is included in the rating: "The author/editor states that this test is not intended for predictive use. Criterion validity is thus not applicable. However, when this test is employed in situations where prediction is indeed a factor, the rating should be 'insufficient', because no research on criterion validity has been performed".

Questions for criterion 7 Criterion validity				
		ins.	suf.	good
Key question 7.1	Is there information about the test-criterion relationship? If the rating of this question is negative (1), skip the rest of the questions.	1		3
7.2	Are the results sufficient with respect to the intended type of decisions to be based on the test?	1	2	3
7.3.a	Are the procedures for calculating the criterion validity coefficients correct?	1	2	3
7.3.b	Are the samples for calculating the criterion validity coefficients consistent with the intended use of the test?	1	2	3
7.3.c	What is the quality of the criterion measures?	1	2	3
7.3.d	Is the quality of the research, as rated in questions 7.3.a through 7.3.c, good enough to confirm the rating of the criterion validity as given in question 7.2?	1	2	3

Notes for key question 7.1: 'Is there information about the test-criterion relationship?'

Points for investigation include for example:

- The correlation of scores on an intelligence test with school performance.
- The predictive value of a test used for the selection of job applicants (for example, validity coefficients or success ratios).
- When making a clinical diagnosis: data on the sensitivity (the ratio between the number of persons identified by the test and the actual number of persons with the disorder) and specificity (the ratio between the number of persons identified by the test as not having the disorder and the actual number of persons without that disorder), and/or data on the ROC curve.

This kind of data does not need to be collected again for each new test in each new situation. The principle of validity generalisation can be used. In this case, the size of the validity coefficients in the original study must be rated in question 7.2, while the quality of the original research study must be rated using question 7.3.

Notes for question 7.2: 'Are the results sufficient with respect to the intended type of decisions to be based on the test?'

Whether one or more validity coefficients suffice depends on a number of factors. Key elements include the purpose of the test, the size of the validity coefficients or the values of the ROC curves, the confidence intervals of these coefficients, the value of the test compared to all other sources of information, the selection ratio and the utility. Furthermore, a test can produce different coefficients in varying situations and groups, or the test may predict some criterion components better than others. Accordingly, in selection situations a validity coefficient of .40 is considered good (see Schmidt & Hunter, 1998), whereas higher coefficients are easily obtained in educational settings. Swets (1988) presents an overview of ROC curve values that have been found in different areas. For certain types of medical diagnosis, these prove to fall between .81 and .97, while for lie detection they lie between .70 and .95. For prediction of school performance (pass/fail) with capacity tests, values between .91 and .94 have been found. The more explicit the author is about the purpose of the test, the better the reviewer can judge whether the test's contribution is effective. It should thus be left to the reviewer's expertise to make a judgement on the size of the values obtained.

The author must sometimes do research on possible prediction bias for the groups in question. This would be called for when the mean scores of subgroups vary, or research on comparable tests reveals that the predictive value can vary among subgroups.

Notes for question 7.3.a: 'Are the procedures for calculating the criterion validity coefficients correct?'

Some aspects which must be considered are:

- Does criterion contamination play a role? That is to say, are predictor and criterion scores established independently? For instance, this is not the case when the supervisor who rates the criterion knows the results of the test.
- Is the time interval between test administration and criterion measurement consistent with the intended use of the test? In validity research, concurrent validity research is quite often resorted to because follow-up data are not available, or the researcher does not want to wait. In the context of selection, this is called the 'present employee method' (Guion, 1991). In principle, the validity coefficients obtained in this manner are less suitable, because it is quite unclear whether they give a proper estimate of the true validity of the test. This is because the composition (selection, dropout), knowledge (experience), motivation and filling-in behaviour (faking) of the research samples during a predictive and a simultaneous study can show differences. In fact the effects of these factors appear to cancel each other out more or less in the selection situation, which means that meta-analyses show hardly any difference in the size of validity coefficients from predictive and concurrent research. Nonetheless, caution should be exercised when interpreting the outcomes of individual studies.
- Did the validity research take place under the same conditions as those in which the test will be used?
- When corrections for attenuation or for restriction of range have been made, are the uncorrected coefficients and other relevant information also mentioned? In certain cases, these corrections produce under- or overestimates of the validity coefficient. After the test has passed its developmental stage, under no circumstances may the correction for attenuation be applied for unreliability in the test itself. In practice, after all, one uses the test score for prediction, and not the true score.
- Has cross-validation research been performed? This is especially important for limited group sizes and certain analysis techniques that capitalise on chance to a large extent; these are primarily multivariate methods like logistical regression analysis and discriminant analysis.
- Is the size of the sample given? The smaller the sample, the larger the confidence intervals of the regression weights and validity coefficients will be.
- If validity generalisation is used, the test author will have to make it plausible that the situations or the tests for which generalisation is claimed are similar. For the similarity of tests, the author will have to show that the same construct is measured with at least equal reliability. This can be especially important for a Dutch translation of a foreign test for which a great deal of foreign research data is already available. If the test author wants to use these data to support the validity of the Dutch version, he will first have to use a technique such as confirmative analysis to prove the equivalence of both versions. If the outcome is positive, the validity coefficients

may be included in the review. This is only possible if these data are summarised in the Dutch manual in an adequate manner.

Notes for question 7.3.b: 'Are the samples for calculating the criterion validity coefficients consistent with the intended use of the test?'

Research on criterion validity must be related to the population for which the test is used. The primary issue here is the variance in the test scores of the sample. It is known that validity coefficients obtained in a heterogeneous group cannot be generalised to a homogeneous group because validity coefficients always decrease dramatically when a homogeneous group is used. It is therefore incorrect to validate a test intended for therapy selection among people who voluntarily requested it (homogeneous group) on a cross-section of the general public (heterogeneous group) because this will produce biased results. To judge this, the sample must be described with the use of any relevant psychological or demographic variables.

Notes for question 7.3.c: 'What is the quality of the criterion measures?'

Sometimes the choice of a criterion is obvious and easily available (passing/failing, a number grade). In other cases, criterion measures have to be separately constructed and collected. In both cases, the criterion must be described as completely as possible, and it must be indicated what relevant behavioural aspects are included in the criterion measure and which are not. In this process, one must consider both construct underrepresentation (not all relevant aspects of the criterion are measured) and construct overrepresentation (some aspects are measured that are not related to the criterion).

Wherever possible, the reliability of the criterion measure should be stated. This is especially true for composite criteria. When the inter-correlations of the separate components of a criterion are low, it is better to state separate validity coefficients for each of the components.

Notes for question 7.3.d: 'Is the quality of the research, as rated in questions 7.3.a through 7.3.c, good enough to confirm the rating of the criterion validity as given in question 7.2?'

Negative answers to one of the questions 7.3.a through 7.3.c results in a rating of '1' for question 7.3.d. This means that the rating for the results of the construct validity research as given in 7.2 must be adjusted downwards. Multiple ratings of '2' on the questions 7.3.a through 7.3.c can also mean that the research is so flawed that question 7.3.d receives a negative rating, and that on this basis the rating for question 7.2 must be adjusted downwards.

Rules for determining final rating for criterion 7 Criterion validity			
The key question is rated '3'.	Question 7.2 is rated '3'.	Question 7.3.d is rated '3'.	good
		Question 7.3.d is rated '2'.	sufficient
		Question 7.3.d is rated '1'.	insufficient
	Question 7.2 is rated '2'.	Question 7.3.d is rated '3'.	sufficient
		Question 7.3.d is rated '2'.	sufficient
		Question 7.3.d is rated '1'.	insufficient
	Question 7.2 is rated '1'.		insufficient
The key question is rated '1'.			insufficient

Literature

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

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

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, 37, 1-15.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education.

Barelds, D. P. H., Luteijn, F., Dijk, H. van, & Starren, H. (2007). NPV-2. *Nederlandse Persoonlijkheidsvragenlijst-2 (Dutch Personality Questionnaire 2)*. Amsterdam: Harcourt Test Publishers.

Bartram, D. (2005). Computer-Based Testing and the Internet. In A. Evers, N. Anderson & O. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 399-418). Oxford, UK: Blackwell.

Bechger, T., Hemker, B., & Maris, G. (2009). *Over het gebruik van continue normering (On the Use of Continuous Norming)*. Arnhem: Cito.

Berge, J. M. F. ten, & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, 69, 613-625.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, 56, 137-172.

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

Campbell, D. P. (1971). *Handbook for the Strong Vocational Interest Blank*. Stanford, CA: Stanford University Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

Cascio, W. F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle River, NJ: Pearson Prentice-Hall.

Cizek G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice*, 15, 13-21.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cohen, J. (1969). Rc: a profile similarity coefficient invariant over variable reflection. *Psychological Bulletin*, 71, 281-284.

Cook, M. (2004). Personnel selection. *Adding value through people* (4th edition). Chicester, UK: John Wiley & Sons.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.

Drenth, P. J. D., & Sijtsma, K. (1990). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen (Test Theory: Introduction to the Theory of Psychological Testing and its Applications)*. Houten/Antwerpen: Bohn Stafleu Van Loghum.

Drenth, P. J. D., & Sijtsma, K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen* (4e herziene druk) *(Test Theory: Introduction to the Theory of Psychological Testing and its Applications, 4th revised edition)*. Houten: Bohn Stafleu van Loghum.

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

Erkens, T. T. M. G., & Moelands, H. A. (1992). *Toetsen met open vragen: een handleiding voor het construeren van toetsen met open vragen (Tests with Open-ended Questions: A Handbook for the Construction of Tests with Open-ended Questions)*. Arnhem: Cito.

Evers, A. (1979). *Amsterdamse Beroepen Interessen Vragenlijst. ABIV. Manual (Amsterdam Vocational Interest Questionnaire)*. Lisse: Swets & Zeitlinger.

Evers, A. (1992). *Amsterdamse Beroepen Interessen Vragenlijst. ABIV92. Manual. (Amsterdam Vocational Interest Questionnaire)*. Lisse: Swets Test Services.

Evers, A., Caminada, H., Koning, R., Laak, J. ter, Maesen de Sombreff, P. van der, & Starren, J. (1988). *Richtlijnen voor ontwikkeling en gebruik van psychologische tests en studietoetsen (Guidelines for the Development and Use of Psychological and Educational Tests)*. Amsterdam: NIP.

Evers, A., & Resing, W. C. M. (2007). *Het drijfzand van didactische leeftijdsequivalenten (The quicksand of didactic age equivalents). De Psycholoog (The Psychologist)*, 42, 466-472.

Evers, A., Vliet-Mulder, J.C. van, & Groot, C. (2000). *Documentatie van Tests en Testresearch in Nederland, dl. 1 en 2 (Documentation of Tests and Test Research in the Netherlands, Parts 1 and 2)*. Amsterdam/Assen: NIP/Van Gorcum.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27, 857-871.

Green, S. A., & Yang, Y. (in press). Coefficient alpha: a cautionary tale. *Psychometrika*.

Guion, R. M. (1991). Personnel assessment, selection, and placement. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial & Organizational Psychology* (Vol. 2, 2nd ed., pp. 327-397). Palo Alto, CA: Consulting Psychologists Press.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255-282.

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

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

Heuvelmans, A. P. J. M., & Sanders, P. F. (1993). Beoordelaarsovereenstemming (Inter-rater agreement). In T. J. H. M. Eggen & P. F. Sanders (Red.), *Psychometrie in de praktijk (Psychometrics in Practice)*, pp. 443-470. Arnhem: CITO.

Hofstee, W. K. B., Campbell, W. H., Eppink, A., Evers, A., Joe, R. C., Koppel, J. M. H. van de, Zweers, H., Choenni, C. E. S., & Zwan, T. J. van der (1990). *Toepasbaarheid van psychologische tests bij allochtonen (Applicability of Psychological Tests to Ethnic minorities)*. LBR-reeks nr.11. (LBR series No. 11). Utrecht: LBR.

International Test Commission (2000). *ITC Test Adaptation Guidelines*. www.intestcom.org.

Kersting, M. (2006). *"DIN SCREEN". Leitfaden zur Kontrolle und Optimierung der Qualität vor Verfahren und deren Einsatz bei beruflichen Eignungsbeurteilungen (Guidelines for Control and Optimisation of Testing Techniques and their Application in Occupational Assessments)*. Lengerich: Pabst Science Publishers.

Keuning, J. (2004). *De ontwikkeling van een beoordelingssysteem voor het beoordelen van "Computer Based Tests" (The Development of a Review System for the Evaluation of Computer-based Tests)*. POK Memorandum 2004-1. Citogroep: Arnhem.

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

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content sensitive item selection in computerized adaptive tests. *Applied Measurement in Education*, 4, 241-261.

Laak, J. J. F. ter, & Goede, M. P. M. de (2003). *Psychologische diagnostiek. Inhoudelijke en methodologische grondslagen (Psychological Diagnostics: Principles of Content and Methodology)*. Lisse: Swets & Zeitlinger.

Langenbucher, J. W., Labouvie, E., Martin, C. S., Suanjuan, P. M., Bavly, L., & Kirisci, L. (2004). An application of item response theory analysis to alcohol, cannabis, and cocaine criteria in DSM-IV. *Journal of Abnormal Psychology*, 113, 72-80.

Laros, J. A., & Tellegen, P. J. (1991). *Construction and validation of the SON-R 51/2-17, the Snijders-Oomen non-verbal intelligence test*. Groningen: Wolters-Noordhoff.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance of educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Luteijn, F., Starren, H. & Dijk, H. van. (1985). *Nederlandse Persoonlijkheidsvragenlijst. Handleiding (herziene uitgave (Dutch Personality Questionnaire: Manual (revised version))*. Lisse: Swets & Zeitlinger.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114, 449-458.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum.

Moelands, H. A., Noijons, J., & Rem, J. (1992). *Toetsen met gesloten vragen: een handleiding voor het construeren van toetsen met meerkeuze vragen. (Tests with Closed Questions: A Handbook for the Construction of Multiple-Choice Tests)*. Arnhem: Cito.

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

Murphy, K. R., & Davidshofer, C. O. (1998). *Psychological testing. Principles and applications*. Upper Saddle River, NJ: Prentice Hall.

Nederlands Instituut van Psychologen (2004). *Algemene Standaard Testgebruik (AST) (General Standard Test Use)*. Amsterdam: NIP.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32, 1-13.

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

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, 21, 173-184.

Resing, W., & Drenth, P. (2007). *Intelligentie. Weten en meten (2e editie) (Defining and Measuring Intelligence, 2nd edition)*. Amsterdam: Uitgeverij Nieuwezijds.

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

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262-274.

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.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17, 277-292.

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

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceeding of the 27th annual meeting of the Military Testing Association (pp. 973-977)*. San Diego, CA: Navy Personnel Research and Development Center.

Twenge, J. M. (2000). The age of anxiety? Birth cohort change in anxiety and neuroticism, 1952-1993. *Journal of Personality and Social Psychology*, 79, 1007-1021.

Veldhuijzen, N. H., Goldebeld, P., & Sanders, P. F. (1993). Klassieke testtheorie en generaliseerbaarheidstheorie (Classical Test Theory and Generalisability Theory). In T. J. H. M. Eggen & P. F. Sanders (Red.), *Psychometrie in de praktijk (Psychometrics in Practice)*, pp. 33-82. Arnhem: CITO.

Verstralen, H. H. F. M. (1993). Schalen, normen en cijfers (Scales, Norms and Numbers). In T. J. H. M. Eggen & P. F. Sanders (Red.), *Psychometrie in de praktijk (Psychometrics in Practice)*, pp. 471-509. Arnhem: CITO.

Vijver, F. van de, & Hambleton, R. K. (1996). Translating tests: some practical guidelines. *European Psychologist*, 1, 89-99.

Visser, R. S. H., Vliet-Mulder, J. C. van, Evers, A., & Laak, J. ter (1982). *Documentatie van tests en testresearch in Nederland (Documentation of Tests and Test Research in the Netherlands)*. Amsterdam: NIP.

Vos, H. J., & Knuver, J. W. M. (2000). Standaarden in onderwijsevaluatie (Standards in educational evaluation). In R. J. Bosker (Ed.), *Onderwijskundig lexicon (Editie III), Evalueren in het onderwijs (Educational Lexicon, Edition III, Evaluation in Education)*, pp. 59-76. Alphen aan de Rijn: Samsom.

Wools, S., Sanders, P., & Roelofs, E. (2007). *Beoordelingsinstrument: Kwaliteit van competentie assessment (Review Instrument: Quality of Competence Assessment)*. Cito. Arnhem.

Zeeuw, J. de (1978). *Algemene psychodiagnostiek II. Testtheorie (General Psychodiagnostics II. Test Theory)*. Amsterdam: Swets & Zeitlinger.

Zegers, F. E. (1989). Het meten van overeenstemming (Measuring Agreement). *Nederlands Tijdschrift voor de Psychologie (Dutch Journal of Psychology)*, 44, 145-156.

Zegers, F. E., & Ten Berge, J. M. F. (1985). A family of association coefficients for metric scales. *Psychometrika*, 50, 17-24.

COTAN/NIP

www.psynip.nl

P.O box 2085
3500 GB UTRECHT
Phone number (030) 820 15 00

This translation of the COTAN Review System for Evaluating Test Quality is drafted by Vertaalservice Radboud in'to Languages and is edited on behalf of the COTAN by Arne Evers, Wouter Lucassen and Karin Vermeulen.