Running Head: DUTCH RATING SYSTEM

The Revised Dutch Rating System for Test Quality

Arne Evers

Work and Organizational Psychology

University of Amsterdam

and

Committee on Testing of the

Dutch Association of Psychologists

Abstract

This article describes the 1997 revision of the Dutch Rating System for Test Quality used by the Committee of Test Affairs of the Dutch Association of Psychologists (COTAN) (Evers, 2001). The revised rating system evaluates the quality of a test on seven criteria: Theoretical basis and the soundness of the test development procedure, Quality of the testing materials, Comprehensiveness of the manual, Norms, Reliability, Construct validity, and Criterion validity. For each criterion, a checklist with a number of items is provided. Some items (for each criterion at least one) are so-called key questions, which check whether certain minimum conditions are met. If a key question is rated negative, the rating for that criterion will automatically be 'insufficient'. To enhance a uniform interpretation of the items by the raters and to explain the system to test users and test developers, comment sections provide detailed information on rating and weighting the items. Once the items have been rated, the final grades ('insufficient', 'sufficient' or 'good') for the seven criteria are established by means of weighting rules.

Keywords: Test quality, test ratings, the Netherlands

In this article the English translation of the Dutch rating system for test quality is presented. The current system is a major revision of the system that is in use by the Dutch Committee of Test Affairs since 1982. The history and background of the system as well as a summary of the rating results are given in Evers (2001). The rating system consists of a series of questions with respect to seven theoretical, practical and psychometric criteria of test construction. The rating results are included in the volumes of Documentation of Tests and Testresearch in the Netherlands, which are periodically published by the Dutch Association of Psychologists (e.g. Evers, Van Vliet-Mulder & Groot, 2000). These publications are a major instrument in the dissemination of information about tests and test quality in the Netherlands.

This article consists of two sections. In the first section the questions are reproduced for each of the seven criteria. In the second section the general comments for each criterion, the comments with respect to each item, and the weighting rules for combining the item-ratings (in order to get the rating for each criterion) are described.

I. CRITERIA AND CHECKLISTS

1. Theoretical basis of the test

1.1. Key question: Are the purposes and applications of the test specified? (If the rating of this item is negative, proceed to item 2.1).

1.2. Is the reasoning behind the test described, and is (are) the construct(s) that the test measures clearly defined?

1.3. Is the relevance of the test content for the construct to be measured justified?

2. The quality of the test materials, and the test manual

2A. Test materials

2.1. Key question: Are the test items standardized?

2.2.a. Key question: Is there an objective scoring system? or:

2.2.b. Key question: Is there a clearly and complete specified system for rating or observation in case the test has to be scored by raters or observers?

2.3. Key question: Are the test items free from racist content or any other offensive content for specific groups of people? (If the rating of one of the items above is negative, proceed to item 2.7).

2.4.a. Are the items, test booklets, answering scales, and answer sheets devised in such a way that filling-in errors can be avoided?

2.4.b. How is the quality of the test materials?

2.5. Is the scoring system devised in such a way that errors can be avoided?

2.6. Are the instructions for the test taker complete and clear?

2B. Test manual

2.7. Key question: Is a test manual available? (If the rating of this item is negative, proceed to item 3.1).

2.8. Are the instructions for the test administrator complete and clear?

2.9. Is specified in what kind of situations the test may be used, how it should be used, and what the limitations of the test are?

2.10. Is a summary of the research published in the manual?

2.11. Is illustrated (for instance by giving some case descriptions) how the test scores may be interpreted?

2.12. Is indicated what kind of information may be important for the interpretation of the test scores?

2.13. Is specified which professional qualifications are required to administer and to interpret the test?

3. Norms

3.1. Key question: Are norms provided (including expectancy tables and cut-off scores)?

3.2. Key question: What is the quality of the normative data and the design of the norming study? (If the rating of one of the items above is negative, proceed to item 4.1).

3.3. Is indicated to which other groups the norms can be generalized; and if so, are the margins of error indicated?

3.4. Are the meaning and the limitations of the type of scale or scores that is used explained, and does the scale prove to be consistent with the intended purpose of the test?

3.5. Is there information about the means, the standard deviations, and the score distributions?

3.6. Is there information about the standard errors of measurement, and/or the standard errors of estimate with the appropriate confidence intervals?

3.7. Is there information about possible differences between subgroups (for instance with respect to gender and ethnicity)?

3.8. Is the year of data collection for each norm group reported?

4. Reliability

4.1. Key question: Is there information about the reliability of the test? (If the rating of this item is negative, proceed to item 5.1).

4.2. Are the outcomes of the reliability research sufficient with respect to the type of decisions that are based on the test?

a. Parallel-form reliability

b. Internal consistency reliability

c. Test-retest reliability

d. Inter-rater reliability

4.3. What is the quality of the reliability research?

a. Are the procedures for computing the reliability coefficients correct?

b. Are the samples for computing the reliability coefficients consistent with the intended use of the test?

c. Is it possible to make a thorough judgment of the reliability of the test on the basis of the information given?

5. Validity

5A. Construct validity

5.1. Key question: Is there information about the construct validity of the test? (If the rating of this item is negative, proceed to item 5.4).

5.2. Do the outcomes of the validity research support the intended meaning of the construct (or: Do the outcomes of the validity research make clear what is being measured)?

5.3. What is the quality of the construct validity research?

a. Are the procedures used in obtaining and computing data on the construct validity correct?

b. Are the samples used in the research on construct validity consistent with the intended use of the test?

c. What is the quality of the other measures used in the construct validity research?

d. Is it possible to make a thorough judgment of the construct validity of the test on the basis of the information given?

<u>5B. Criterion validity</u>

5.4. <u>Key question:</u> Is there information about the relationship test-criterion? (If the rating of this item is negative, items 5.5 and 5.6 can be skipped).

5.5. Are the outcomes of the validity research sufficient with respect to the type of decisions for which the test is intended?

5.6. What is the quality of the criterion validity research?

a. Are the procedures used in obtaining and computing data on the criterion validity correct?

b. Are the samples used in the research on criterion validity consistent with the intended use of the test?

c. What is the quality of the criterion measures?

d. Is it possible to make a thorough judgment of the criterion validity of the test on the basis of the information given?

II. COMMENTS AND WEIGHTING RULES

1. Theoretical basis of the test

General comments

Test construction demands thorough preparation. After all, test scores often serve as the basis for making sound judgements with respect to intra-individual differences (within such contexts as educational follow-up systems and vocational guidance), inter-individual differences (for example in personnel selection), and differences between groups or situations (as in organizational diagnosis). The information provided by the test author should enable the prospective test user to judge whether the test is suitable for his or her purposes. With this in mind, the first step should be to provide a clear description of the construct the test intends to measure. The choice of the test content and the method by which a construct is measured should also be accounted for.

This category is concerned only with the viability of the theory and the comprehensiveness of the description of the basic assumptions. The question of whether the test developer has been successful in constructing a reliable and valid test is considered in Criteria 3, 4, and 5, which focus on the quality and the results of the research.

[insert Table 1 about here]

Recommendations per item

Item 1.1. The development of a test should start with a reflection on the purpose of the test. Is the aim to predict some type of criterion behaviour? Is a test intended to evaluate educational progress or the effects of training? Is it being used to assess suitability for a particular type of treatment?

It is also essential to specify the group or groups for whom the test is intended, detailing such aspects as age group, occupation, educational or occupational level, and whether the context in question is normal or clinical. However, the more ambitious the test author's claims, the greater his or her obligation to deliver empirical data, such as norms and evidence of validity.

Item 1.2. The test author should make clear whether the test reflects an existing theory or whether it is based on newly developed ideas which may eventually bring about major or minor changes in this theory. Whatever theory the test is based on, it must be described in sufficient detail.

If the test is a translation or an adaptation of a foreign instrument, sufficient background information 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 make clear which facets of behaviour belong to the domain.

When a traditional or historically based test method (as opposed to one with a theoretical foundation) is used to measure a construct, arguments must be given to show why it is useful to measure these constructs. Similarities and differences with similar tests must also be described. From this description, the added value of the new instrument over existing instruments should be evident.

Item 1.3. This question applies to the steps from the construct to be measured through to operationalization. The definition of the content domain must make it possible to judge which items belong to the domain and which do not. This could be accomplished by analysing the construct in such a way that it becomes clear which facets can be distinguished. Theoretical or content-related considerations for the weighting given to these facets should also be stated, and the question of whether item-sampling takes such weighting into account should also be addressed. Whenever items are dropped or modified in the course of constructing or adapting a test, the consequences of these changes for the measurement of the original construct must be indicated (the content domain may be shifted, narrowed or incompletely covered as a result of these changes).

2. The quality of the test materials, and the test manual

General comments

A test score can only be interpreted as a reliable measure if the test is administered under standardized conditions. The purpose of standardizing the test situation is to prevent unintended factors from affecting the test score. For example, administration and instruction should be standardized to such an extent that the influence of a different test administrator or variation in instruction is eliminated or minimized. Criterion 2 is divided into two sections: the first consists of a group of questions relating to the quality of the test materials and the second concerns the quality of the manual. A final rating is given for each of these subsections separately.

The questions on the quality of the test materials refer to the design, the content, and the form of the test materials, the instructions for the test taker, and scoring. This section has three key questions. A rating of ‘+/-’ for the third key question may mean that the final rating is accompanied by an additional comment, such as ‘Of very restricted usability for ethnic minorities’. The second section of this chapter deals with the completeness of the information provided in the manual with respect to the possible uses of the test and the interpretation of the test scores.

[insert Tables 2A and 2B about here]

Recommendations per item (Criterion 2A)

Item 2.1. Test items are standardized when they are the same for every respondent with respect to content, form and order. Standardization is an important condition for interpreting and comparing scores. An exception with respect to the requirement of an uniform order of test items is made for adaptive tests. However, for this type of test, the rules for deciding the choice of any subsequent item must be made explicit.

Item 2.2.a. A scoring system is called objective when the score values accorded to all possible answers for all the items in a test are such that any qualified person who scores the items will, apart from clerical errors, give exactly the same score to the same answers. This is particularly applicable to paper-and-pencil ability tests and questionnaires with multiple choice items.

Item 2.2.b. For observation scales, projective tests, subtests of individual intelligence tests, and essay questions, scoring cannot be strictly objective.

Item 2.3. In 1990 a task force jointly established by the Dutch Association of Psychologists and a national anti-discrimination body screened 20 of the most frequently used Dutch tests for racist and ethnocentric content. The content of a test was qualified as racist if it could reasonably be expected that words, sentences, pictures or other elements used in items or instructions would be offensive to members of one or more ethnic minority groups. The content of a test was qualified as ethnocentric when the test contained unnecessarily difficult or unnecessarily culture-bound words, idioms or pictures. The task force stated that racist content would make the test unusable, while ethnocentricity would lead to restricted usability. None of the 20 tests were found to have racist content. However, ethnocentricity proved to be quite a common phenomenon (Hofstee et al., 1990). In this rating system, a similar strategy will be adopted: racist content will result in a negative rating on this item, and pronounced ethnocentricity in a plus/minus.

The principle of restricted usability can be applied to other groups as well. One example is an interest inventory containing pictures with markedly gender-stereotyped items. This principle does not apply to tests which are designed to measure concepts related to racism or gender (such as Adorno's F-scale or a scale for androgyny).

Item 2.4.a. Issues which must be taken into account in rating this item include: (a) the comprehensibility of items or assignments for the groups for whom the test is intended (the majority of the items must not go beyond the limitations of these groups); and (b) answer sheets, if used, should be devised in such a way that errors in filling out (such as skipping an item) can be detected quickly.

Item 2.4.b. This item deals with all kinds of practical aspects of the test material which are not covered in the other items for Criterion 2A. These include the following considerations:

- Are the items formulated in everyday language?
- Is the test legible?
- Can colours or symbols (if applicable) be clearly distinguished from one another (even for colour-blind individuals)?
- Are the test materials durable?

Item 2.5. In rating this item, attention has to be paid to points such as the following:

- The scoring procedure has to be described completely and clearly.
- If hand-scoring keys are being used, there must be clear instructions on how to place these keys on top of the answer sheets.
- If hand-scoring keys are being used, the keys have to fit neatly over the answer sheets.
- If hand-scoring keys are being used, the version of the test has to be printed on the keys. This is particularly important when a test is revised.
- There should be clear instructions on the scoring of missed items.
- An indication must be given as to how many items can be missed without the test losing its value.
- If raters or observers are involved, there should be clear indications on how to deal with differences between raters or observers.

In general the use of separate answer sheets is preferable, because the scoring of answer sheets will result in less errors than scoring a test booklet. If the administration and scoring of a test is computerized, the rater of the test must be given an opportunity to check the scoring.

Item 2.6. In this rating system a distinction is made between the instructions for the test taker and those for the test administrator. The quality of the instructions for the test taker are rated in this item; the instructions for the test administrator are rated in item 2.8. The instructions or recommendations for the test taker form part of the test material and usually constitute the first page or pages of the test booklet or the first page or pages of text when the test is administered by computer. The instructions must be standardized and should include the following elements:

- One or more sample questions.
- Information about how to record (or type, in case of computerized testing) the answers.
- A strategy for guessing or for answering when alternatives are of the same degree of likelihood or applicability.
- Time limits.

Recommendations per item (Criterion 2B)

Item 2.7. A test author must provide a manual containing practical information (commonly called a User's Guide) and technical information (commonly called a Technical Manual). Dissertations, journal articles or research papers should not be regarded as a manual.

Item 2.8. The main objective of the recommendations for the test administrator in the manual is to ensure the standardization of the test. The outline should be as explicit as possible in its description of exactly what the test administrator has to say, what he or she must not do or say, and the tasks he or she must perform (such as arranging the test materials in a certain order for an ability test). A recommendation along the lines of "the test administrator explains the purpose of the test to the test taker" would be insufficient. The test manual must also provide a detailed outline of how to deal with questions commonly asked by test takers.

Item 2.9. A manual must be complete, accurate, and clear about the applicability of the test. This may lead to suggestions for using the test differently, depending on the specific situation or application. Examples of points to consider include:

- Has it been made clear 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 the case of progress evaluation?
- Can test results obtained in a clinical situation lead to empirically founded conclusions or can they only serve 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?
- In the case of tests for personnel selection, have the job classes for which the test is intended been indicated and has the critical content of these jobs been specified?

Item 2.10. For test users and prospective test users the manual will be the principal source of information, because many of them will not have easy access to dissertations, research reports or other published materials, and because they will not have the time for (or be sufficiently interested in) reading all the technical details. A summary of norming, reliability and validity studies must therefore be provided in the manual. If new research provides useful additional information, users should be informed by means of supplements to or revisions of the manual. In this item only the availability of the information in the manual is assessed. The quality of the research designs and results are evaluated in the Criteria 3, 4 and 5. If this kind of information is not published in the manual this item will be given a negative rating. A negative rating has no implications for the rating of Criteria 3, 4 and 5, because the rater will always consult the original publications in order to make a sound judgment of the psychometric qualities of the test.

Item 2.11. Generally speaking, the inclusion of a number of case descriptions in the manual should help the test user with the interpretation of test scores.

Item 2.12. Has an indication been given as to the other variables which contribute to the prediction?

Item 2.13. Statements on user qualifications should detail the specific training, certification, or experience needed.

3. Norms

General comments

Scoring a test usually results in a so-called raw score. Taken by itself, the raw score has little or no significance. In general, the raw score can be understood by referring to a norm. The norm may be derived from a domain of skills or subject matter to be mastered (content-referenced interpretation), or may be based on the distribution of scores from a reference group (norm-referenced interpretation).

In content-referenced interpretation the results of each test taker are examined separately to see which answers were right and which were wrong. The results are not compared with the results of others. However in norm-referenced interpretation, comparison with others is the essence of the process. In norm-referenced interpretation the score of a test taker is compared with the scores of other individuals with whom a useful comparison can be made (on the basis of similarities in age, grade, job etc.).

A norm of one type or the other is a basic requirement of all tests. However, there can be exceptions, especially in the case of ipsative tests, where only intra-individual interpretation is recommended. In such cases, the questions in this chapter need not be answered, and the qualification ‘not applicable’ can be used.

Norms are susceptible to erosion. Of all the psychometric properties of a test, norms are the most sensitive to such factors as social changes, educational changes and changes in job content. Consequently a test either has to be renormed from time to time, or the test author has to show by means of research that there is no need for renorming. To draw the attention of the test user to the possibility that norms may be out of date, the qualification “The norms are outdated” is added to the rating of tests if the norm data were collected more than 15 years previously. After another five years without renorming have passed, this qualification is changed to: “The norms are unusable because they are outdated”. These qualifications are published once a year.

[insert Table 3 about here]

Recommendations per item

Item 3.1. Norms must be available at the time the test is published for actual use. Normative data based on individual scores are not relevant to tests intended for norm-referenced assessment of groups and vice versa. Norms can no longer be used when changes have been made to a test, for instance in the event of modifications to items or instructions. The conversion of a paper-and-pencil test to a computerized version is usually of little influence on the scores of questionnaires, but in the case of ability tests, usually new norming data must be collected. This is especially important where time limits are involved.

Item 3.2. Basically norms have to be presented for all purposes and applications for which the test is recommended by the test author (see item 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 selections for technical jobs, 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 (that is to supply a reliable set of reference points): Firstly, it must be representative of the referred group, and secondly the norm group has to be sufficiently large. To facilitate an assessment of the representativeness of the norm groups, a definition of the population and a complete description of the sampling design and the process of data collection must be given. It is often the case that the information given is too limited. It must be evident from the description which population is concerned, whether data were collected locally or nationally, whether the data collection covered an average of the population or only targeted individuals with specific qualities (for instance people with mental problems or with a specific educational background), what the participation rates were and so on. Quite often data are collected by means of a so-called ‘sample of convenience’. Such a sample might consist of the clients of a vocational guidance service, psychology students or the patients from a specific psychiatric hospital, for example. In general samples of this kind make poor norm groups because the reason for testing these subjects can be related to the test score. These kinds of samples cannot therefore be considered representative for the intended population (which, for the examples given above, could be high-school students, college students and people with psychiatric problems respectively).

Recommendations on the desired sample size are scarce in the literature (Angoff, 1971; Campbell, 1971). These recommendations are either based on the calculation of errors of measurement in parameters for score distribution, or on empirical evidence with respect to the stability or instability of these parameters. A combination of the outcomes of these two methods, linked to the importance of the decisions for which the test is intended, has resulted in the following rules for rating the adequacy of sample sizes:

- Tests intended for important[1] decisions at individual level (for instance, personnel selection, placement in special educational programmes, admission for/discontinuation of clinical treatment):

N < 300 is INSUFFICIENT; 300 ≤ N < 400 is SUFFICIENT; N ≥ 400 is GOOD.

- Tests intended for less important decisions at individual level (for instance assessment for learning, and general descriptive use of test scores in such areas as vocational guidance and evaluation of treatment outcomes):

N < 200 is INSUFFICIENT; 200 ≤ N < 300 is SUFFICIENT; N ≥ 300 is GOOD.

- Tests intended for research at group level:

N < 100 is INSUFFICIENT; 100 ≤ N < 200 is SUFFICIENT; N ≥ 200 is GOOD.

Though the requirement with respect to sample size applies for each norm group, there can be exceptions. For instance, when a developmental test is normed for different age groups separately, the sample size of each norm group is a significant factor. However, if fit procedures are applied (using the data of all age groups simultaneously) the sample size of the combined groups is more relevant.

[1] Important decisions are: decisions taken on basis of the test scores that are essentially, or in the short term, irreversible, and on which the test taker has little influence.

Item 3.3. The collection of norming data is a costly and labour-intensive process that need not be carried out for all possible groups. Norms for intermediate age or year groups can be acquired by means of extrapolation or norms can be generalized to similar jobs or occupations. The test author must justify the extrapolation or generalization of norms by describing the critical points of similarity between groups (for instance in terms of job content or education) or by providing research data (such as scores on related variables).

When age or grade norms are provided, the width of the age or grade interval contributes to bias in the test scores: performance is underestimated at the beginning of an interval and overestimated 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 prevented by extending the number of norm tables. If a test is intended to be used at a specific time of year, this must be clearly mentioned, and the norming data must be collected within this period. In any case, the mean age in months or the time of year in which the norms are collected have to be mentioned for age-related and educational norms.

Item 3.4. For the transformation of raw scores to derived scores, there is a wide range of scale systems from which to choose. A choice can be made between standard scores and percentiles and a scale with many units or few units. For particular situations, the test author can decide to design a new system or adapt an existing one (Verstralen, 1993). The choice for a specific scale system must be consistent with the purpose of the test, and with the expertise of the test user. When the purpose of the test falls under the category “important” (see recommendation 3.2) it is appropriate to choose a precise system with many units. However, for tests in this category the use of confidence intervals is recommended (see 3.6), which in turn requires high professional qualifications from the test user. Opting for a rough system means sacrificing precision, but may make the results more readily comprehensible. A rough system is preferable when only a broad indication is required, and especially in cases where the level of expertise of the prospective test user is not high. Whatever scale system is used, the features and the possible pros and cons of the system should be described and the reasons for choosing the scale score should be given.

When raw scores are transformed in standard scores, normalizing transformations should be used, unless the raw scores already approximate a normal distribution. Furthermore, because transformations which use the observed cumulative proportions are highly sensitive to sample fluctuations, it is better to apply fit procedures to the cumulative distribution before proceeding any further (Laros & Tellegen, 1991).

When content-referenced interpretation is recommended for a test, the test developer’s rationale for suggesting a given cutting score or presenting certain levels of ability should be accounted for. A thorough description of the relevant domain is essential to such a justification. The quality of this description has been rated in Criterion 1. This item is purely concerned with the justification of the score which makes the difference between pass or fail, admission or rejection.

Item 3.5. These data have to be provided for each norm group. Aspects 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. Another example might be the bottom or ceiling effects in a test for cognitive abilities reflected in the score distributions of groups with a relatively high or low standard of education. A test user needs this kind of information to be able to interpret test scores correctly.

Item 3.6. In the literature (for example Drenth & Sijtsma, 1990; Nunnally & Bernstein, 1994) the standard error of measurement and the standard error of estimate are not differentiated even though they should be (Laros & Tellegen, 1991; Lord & Novick, 1968). The standard error of measurement gives a confidence interval that is symmetrical around the observed score. It indicates whether the observed score differs significantly from each true score that lies within this interval. The formula for the standard error of measurement is $\sigma_{est} = \sigma_x \surd\ (1\text{-}r_{xx})$. The confidence interval is important when the use of significance tests is called for (does the score of person A differ from the score of person B or from a cutting score X_0?).

The standard error of estimate results in a probability interval for the true score. This interval is symmetrical around the true score. Notice that ‘the’ estimated true score is the average true score of all persons in the norm population with the same observed score. The width of the probability interval is based on the deviation of the true scores of these persons. The probability interval is important when estimating the level on the measured variable while taking into account the reliability of the test. Because probability intervals are based on the regression method, the term standard error of estimate is used as it would be in the case of prediction. The formulas for the estimated true score (T’) and the standard error of estimate (σ_{est}) are:

$T' = \mu_x + r_{xx}(X - \mu_x)$ and $\sigma_{est} = \sigma_x \surd\ r_{xx} \surd\ (1\text{-}r_{xx})$.

The formula for the estimated true score is frequently applied, however the interval often is erroneously based on the standard error of measurement. As the formula shows, the standard error of estimate is a factor $\sqrt{r_{xx}}$ smaller than the standard error of measurement.

For norm scores the probability interval has to be computed differently. A complicating factor with norm scores is that the observed scores are standardized and relate to a particular position in the population, aspects which not apply to true scores. The spread of the true scores becomes smaller as the reliability increases. It would therefore be better to construct norm tables after having standardized the distribution of true scores (Laros & Tellegen, 1991). If this has not be done (as is usually the case), a probability interval can nevertheless be computed for the true scores by equating the spread of the true scores with the spread of the normed scores. The formulas for T' and σ_{est} then become:

$T' = \mu_x + \sqrt{r_{xx}}\,(X - \mu_x)$ and $\sigma_{est} = \sigma_x \sqrt{(1-r_{xx})}$.

In these formulas the parameters μ_x, X, and σ_x refer to norm scores. Compared to the previous formulas there is less regression to the mean and the standard error of estimate is equal to the standard error of measurement as it is usually computed.

For the confidence intervals, it is sufficient to give the standard error of measurement and to indicate how the standard error of measurement should be applied. The regression to the mean makes the computation of probability intervals more difficult, which means it is better to incorporate these intervals in the norm tables.

This subject is explored in such depth here because the correctness of the formulas and the interpretation given in the manual are of such importance to the evaluation of this item. For tests in the category 'important', probability intervals have to be provided in order to obtain a positive rating on this item. If only the standard error of measurement is given (correctly computed and with the correct interpretation), a rating of ‘+/-’ will be given. For less important tests, reporting the correct standard error of measurement is sufficient to obtain a positive rating.

Item 3.7. There are various reasons why subgroup differences must be studied and reported:

- The results may show adverse impact.
- The results may justify test- or item-bias research.
- The test user receives data which allow him or her to decide whether or not to account for possible differences.

These studies only apply to those subgroups which are relevant to the purpose of the test, such as groups based on gender, age or ethnic background.

Item 3.8. Stating the year in which the norm data were collected is important for deciding whether norms are outdated and, if so, to what extent.

4. Reliability

General comments

The variance in test scores consists of true and error variance. The sources of error variance can be different. The various reliability indices which can be computed reflect these sources to different degrees. It is therefore not possible to speak of the definitive reliability of a test: different forms of reliability are distinguished depending on the source of error analysed in a particular study. What is more, the results of the reliability study will differ depending on the characteristics of the group studied (the homogeneity with respect to the construct measured has a particularly strong influence on the computed coefficients). In practice, the traditional reliability indices, as mentioned in item 4.2, indicate the extent to which test scores can be generalized with respect to version, items, time and raters. The reliability of a test is generally studied using traditional methods and the items in this chapter have been worded accordingly. However it is also possible to make use of other models, for instance by using analysis of variance to estimate the influence of several facets at a time.

In general only one reliability rating is given, though a test may produce several scores. For example, this is the case for personality and interest questionnaires which contain several scales and for multiple aptitude batteries with several subtests (whether these can be administered independently or not). In such cases the lowest coefficient is decisive for the rating. However, when this coefficient is a clear negative exception among the reliabilities of the other scales or subtests, the higher rating may be given (for instance if the reliabilities of all subtests are rated ‘good’ with the exception of one subtest rated ‘insufficient’). The exception should be mentioned in a footnote.

When the scores on the subtests are added together to arrive at a total score, as is the case with some intelligence tests, there are three possible approaches:

- Only the interpretation of the total score matters. In this case only the reliability of the total score needs to be rated.
- The test author states that the total score is indeed the most important, but that the interpretation of subtest scores is also possible. In this case 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 the comments for item 4.2). For example if the total score is categorized 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.

- The test author indicates no difference in importance between subtest scores and total score. In this case the requirements for both kinds of score are the same.

When the ratings given for the reliability coefficients of subtests and total scores differ, this should be mentioned in a footnote added to the rating. It is also important to note that only one rating is given when a test author provides reliability coefficients computed for several groups. The above rule also applies to such cases. The lowest reliability coefficient is decisive for the rating except in cases where this value is a clear exception.

[insert Table 4 about here]

Recommendations per item

Item 4.1. Some kind of reliability coefficient will generally be provided, but the results of generalizability studies may also be considered.

Item 4.2. No general statement about the desired level of a reliability coefficient can be made, because the purpose for which the test is used must always be taken into account. Nunnally & Bernstein (1994, p. 265) indicate that a test that is used for important decisions must have a reliability of at least r = .90. Taking this value as a basis, the following rules have been formulated:

- Tests intended for important[2] decisions at individual level (for instance, personnel selection, placement in special educational programmes, admission for/discontinuation of clinical treatment):

  $r < .80$ is insufficient; $.80 \leq r < .90$ is sufficient; $r \geq .90$ is good.

- Tests intended for less important decisions at individual level (for instance, assessment of learning, and general descriptive use of test scores in such areas as vocational guidance and evaluation of treatment outcomes):

  $r < .70$ is insufficient; $.70 \leq r < .80$ is sufficient; $r \geq .80$ is good.

- Tests intended for research at group level:

  $r < .60$ is insufficient; $.60 \leq r < .70$ is sufficient; $r \geq .70$ is good.

When estimating variance components in generalizability studies, the lines between the different qualifications are equivalent to the values above.

For educational settings, Frisbie (1988, p. 29) mentions a value of .50 as a minimum for ‘teacher-made tests’ when the scores on these tests are combined with other information (for instance achievement test scores, observations, grades for assignments) to give a final grade. Such a test can effectively be considered a subtest, since it forms part of a total computed score. In view of the fact that this rating system only applies to the rating of separate tests, this kind of use cannot be accounted for (unless the test author explicitly recommends this kind of use, in

[2] ‘Important decisions’ are 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.

which case the rules for rating subtest and total scores apply). However, this example emphasizes that a test with ‘insufficient’ reliability can still be useful in the diagnostic process, provided that it is used in combination with other information.

Item 4.2.a. The generalizability of different forms of a test (parallel forms, for example) can be determined by means of the correlation between the forms with comparable item content and (ideally) with equal item difficulties, means and variance. The correlation is an estimate of the reliability of both forms. Parallel test reliability can be useful for 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 test reliability. A correction for test length may subsequently be applied.

Item 4.2.b. Generalizability with respect to items (or groups of items) within a test is usually computed by means of Cronbach’s coefficient α. As coefficient α is sensitive to the number of items, it is important to be aware that a large number of items can result in a high reliability coefficient even in cases where inter-item correlations are moderate. It is now regarded as inadvisable to use the once popular split-half coefficients, as the results depend on the arbitrary allocation of items to test halves.

Internal consistency or homogeneity indices are not useful for speed tests or for so-called heterogeneous scales. The same applies to causal indicators (see Nunnally & Bernstein, 1994, for the difference between causal and effect indicators). In such cases one of the other reliability measures may prove useful. For pure speed tests the parallel method or the test-retest method may be used (see also comment 4.2.a). However, many power tests also have a time limit. Especially when a significant percentage of the test-takers have not been able to complete the last test items, internal consistency should not be computed automatically, because it can result in reliability being overestimated. In such cases a reasonable 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), administering these halves within half the testing time, computing the correlation between the halves and then correcting the correlation to take account of the halving of the test length. When speed is not a primary factor (that is, if at least 70% of the test-takers complete the last item) a correction formula for the internal consistency coefficient may be applied (De Zeeuw, 1978). Another possibility would be to compute the internal consistency 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 kind of tests correlations with other variables may replace reliability indices. For causal indicators in particular, a thorough specification of the domain is essential (see Criterion 1).

For all kinds of adaptive or tailored tests internal consistency cannot be used indiscriminately. Some authors wrongly assert that an internal consistency index can be used because a high correlation between the adaptive score and the score for the whole test has been demonstrated by simulation or some other means. 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 using various break-off point rules and by computing the correlations of the acquired scores with a criterion variable.

Item 4.2.c. Generalizability over time is estimated by means of test-retest correlations which are computed for the repeated administration of the test to the same group. The length of the time interval between the two administrations must be stated, as well as any relevant events which took place during the interval. If the interval is long enough, it may serve as an indication of the stability of the test scores.

Item 4.2.d. For observation and rating scales in particular, it is important to know whether scores can be generalized with respect to observers or raters. Indexes that can be used are agreement coefficients such as Cohen’s kappa (Cohen, 1960,1969), Gower’s coefficient (Gower, 1971), the identity coefficient (Zegers & Ten Berge, 1985) or other measures that account for differences in means and variances of ratings (see Zegers, 1989 for an overview). A study of the variance components or the factor structure of the observers’ or raters’ behaviour may be relevant here.

When rating the reported values it is important to take the kind of coefficient used into account. For example, in the literature a distinction is made 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 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 used for the various coefficients mentioned by Zegers (1989) are similar. Finally, it has to be stressed that a high inter-rater reliability is an essential condition for a high test reliability, but that it is not the same as a high test reliability.

Item 4.3.a. For each of the four forms of reliability some points of special attention are given below:

1. When tests have different means or standard deviations or when parallelism is not plausible for other reasons, the computed coefficient cannot be regarded as an index of reliability, but should instead be seen as an index of congruent validity.
2. Test or scale development usually aims to achieve the highest possible internal consistency. This may result in a very specific test content which measures a much narrower construct than intended. In general it is not too difficult to obtain high internal consistency by developing items that are almost identical, but such a scale or test may not be very useful. The fact that a subgroup of items in a test show higher intercorrelations than correlations with the other items, or even the existence of several such subgroups within a scale, does not preclude high internal consistency. On the contrary, if correlations with the other items are moderate such homogeneous subgroups enhance the internal consistency. These relatively high intercorrelations may be due to the fact that these items share unintended variance not common to the other items in the test. This may occur when items are formulated similarly, have a specific word in common and so on. Such unintended variance can contribute to a high internal consistency. Generally speaking, the points mentioned above mean that the test partly measures a construct other than that intended (by narrowing the construct or by introducing unintended variance) in order to achieve a high level of homogeneity. This can be avoided by testing for unidimensionality during the developmental phase of a test, using LISREL analyses for example. The test developer can take action depending on the results of these analyses and the theoretical basis of the test (for instance by replacing or rewriting items, or changing the structure of the test). The addition of ‘the theoretical basis of a test’ is

important here, because the test author may explicitly want to measure a narrow or a multidimensional construct. The usefulness of such a starting point and the results of the research into the dimensionality of the test are rated in Criterion 1 (Theoretical basis) and Criterion 5 (Validity) of this rating system respectively. At this point the rater is only asked to assess the size of the internal consistency coefficient, in the light of the phenomena mentioned above, especially when analyses for unidimensionality have not, or not yet, been performed.

3. No strict standards can be given concerning the optimum length of the test-retest interval. As a rule, a very short interval (up to a few weeks) is not suitable because of the role played by memory. A very long interval (longer than a year) may not be useful either, because external events or experiences may influence the individual’s personality and skills, thereby affecting the score on the retest. The test-retest correlation with long intervals is not so much an index of the reliability of the test, but more a measurement of the stability of the trait measured. However, the intervals mentioned above are rather arbitrary. The age of the group tested, the nature of the test itself, and the purpose of the test always has to be taken into account before determining the appropriate interval. The same applies when rating the appropriateness of the interval chosen. For example, a test that is intended for the long-term prediction should have a relatively long test-retest interval.
4. Observations or ratings must be carried out independently when inter-rater reliability is used to estimate the reliability of the test. This fact should be clear from the description of the research design.

Item 4.3.b. Reliability coefficients must be computed for the groups for whom 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 assessment within this reference group which matters. It is therefore incorrect, and even misleading, to compute reliability coefficients for the total of the groups or for a selection of the upper and lower parts of the score distribution (as happens incidentally). Since the size of the reliability coefficient also depends on the spread of the scores, the computed coefficient will almost certainly be higher when the scores of the total group or of the extremes of the score distribution are used instead of the coefficient for each norm group separately.

Item 4.3.c. Below are some examples of information which must be available to facilitate sound assessment of 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, has mention been made of the percentage of test-takers who have answered each item?
- Have the samples for which the reliability coefficients are computed been described in sufficient detail?
- Has the number of observers or raters included in the computation of a reliability coefficient been mentioned?
- 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 in the same way so that the reliability of the ratings can be generalized to their situation. It should 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 is provided, the reported reliability coefficient can be rated ‘insufficient’ because the quality of the research design cannot be verified. In most cases, enough information will be provided to allow the quality of the reliability research to be assessed. In borderline cases especially (insufficient/sufficient, sufficient/good), inadequate information can be a reason for giving the lower rating.

5. Validity

General comments

Validity is the extent to which a test fulfils its purpose. Can the appropriate conclusions be drawn from the test scores? In the literature many forms of validity are mentioned; Drenth and Sijtsma (1990), for example, mention eight. These distinctions reflect the purpose of the validity research or the specific data-analytical technique used in the validation process. This rating system follows the traditional three-category classification with respect to the purpose of the validity research (APA, 1985; Evers et al., 1988). These categories are: construct validity, criterion validity and content validity. Content validity has already been dealt with in Criterion 1, because it is regarded as part of the developmental process of a test.

Construct validity concerns the answer to the question: ‘What is being measured by the test?’ Does the test measure the intented concept or does it partly or mainly measure something else? Frequently used methods or techniques to provide evidence of construct validity are: factor analysis to demonstrate unidimensionality, comparing the mean scores of groups that are expected to differ, and computing correlations with tests that are supposed to measure the same construct (congruent validity). Essentially, this kind of research is relatively easy to perform and the results can give an initial indication of the evidence of construct validity. However, by themselves none of these indications are enough to guarantee a rating of ‘sufficient’. Only the accumulation of such evidence, more extended research on the structure of the construct, or a well-designed study using a multi-trait-multi-method design can result in a rating of ‘sufficient’ or ‘good’.

Criterion validity demonstrates that test scores are systematically related to one or more outcome criteria. In this context the term prediction is generally used. Prediction can focus on the future (predictive validity), the same moment in time (concurrent validity), or on the past (retrospective validity). It is important to specify the kind of criteria for which relations are expected. This is especially true when a test consists of several subtests or scales. However, demonstrating the validity of all subtests or scales is not essential to obtain a rating of ‘sufficient’ or ‘good’, since a single valid scale can make the instrument very useful. Information relevant to criterion validity can also be considered when assessing construct validity. In fact this information also forms part of the process of construct validation (see for example Anastasi, 1986; Messick, 1988), because the information can help to clarify what is being measured by the test.

[insert Table 5 about here]

Recommendations per item (Criterion 5A)

Item 5.1. The information requested here concerns the internal or external structure of the test. The internal structure can be investigated by determining measures of association between (groups of) items, between items and test, and between subtests. Procedures such as having test-takers think aloud or the examination of items can also be used. The external structure usually is investigated by determining the relationship with other tests (convergent and discriminant validity).

Item 5.2. As mentioned above, construct validity in particular is a matter of the accumulation of research evidence. Construct validation research is never completed. Item bias research for different groups can be desirable. An additional advantage of item bias research is that it provides information concerning the possible multidimensionality of the construct as measured.

Item 5.3.a. As a consequence of the diversity of this kind of research, hardly any general guidelines can be given. Some points of attention are:

- When expected outcomes have not been formulated in advance, research outcomes should be interpreted with great caution. Without clear-cut expectations the interpretation of research outcomes can easily degenerate into ‘fishing’: it is always possible to find interpretable relations of some kind when test scores are correlated with the scores of a great number of other variables (available by chance). Some of the significant correlations will in fact be a matter of chance, and it is not possible to see which are significant and which are not.
- Item-rest correlations, instead of item-test or item-total correlations, must be reported.

Item 5.3.b. See comment on Item 5.6.b.

Item 5.3.c. The reliabilities of the measures used must be made known. It will be obvious that validating the test score with measures with a low reliability (lower than .60) is not useful, because the results will be ambiguous (a low correlation can mean that the test is measuring something different or it may be the result of the low reliability of the other test). Moreover, validating a test with a congruent test is only useful if the validity of the other test has itself been sufficiently investigated.

Item 5.3.d. Information on matters such as sample sizes and a sufficiently detailed description of the analysis techniques must be provided.

Recommendations per item (Criterion 5B)

Item 5.4. Examples of criterion-related evidence of validity include: the correlation of scores on an intelligence test with school performance, the predictive value of a test used for the selection of job applicants, and data on the sensitivity and the specificity of a test for clinical diagnosis. This kind of data does not need to be collected again for each new test in each new situation. The principle of validity generalization can be used. If this applies, the results of the original study have to be assessed with item 5.5, while the quality of the original research design has to be assessed using item 5.6.

Item 5.5. 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, the confidence intervals of the coefficients, the value of the test compared to other instruments or other sources of information, the selection ratio and a cost-benefit analysis. Furthermore, a test can have different validities in different situations and for different groups, or the test may predict some criterion components better than others. Accordingly, in selection situations a validity coefficient of .40 is valued as good, whereas higher coefficients are easily obtained in educational settings. The more explicit the test author is about the purpose of the test, the better the rater can judge whether the test makes a useful contribution to this purpose.

When there is reason to consider the possibility of prediction bias, a test author has to investigate this possibility for the subgroups concerned. Such reasons might be differences in the means of subgroups or evidence of prediction bias in similar tests.

Item 5.6.a. 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 performance of an employee knows the results of the test administered earlier during the selection process.
- Is the time interval between the administration of the test and the appraisal of the criterion consistent with the intended use of the test? Due to the fact that follow-up data are not available or not yet available, concurrent validity research is quite often resorted to. In the context of selection, this is called the ‘present employee method’ (Guion, 1991). One of the disadvantages of this method is that, taking the example of selection, experiences on the job may influence the test score and consequently the validity coefficient. Another disadvantage is that other information, like restriction of range, can no longer be retrieved. Evidence of validity, obtained with this method, can therefore never provide an accurate estimate of the true validity of the test.
- Are the test conditions in the validity study the same as the conditions under which the test is actually used?

- When validity generalization is used, does the test author provide sufficient arguments for the similarity of the situations (or the tests) that are subject of generalization? To demonstrate the similarity of tests in order to generalize validity results, it must be shown that both tests measure similar constructs and have similar reliabilities.

<u>Item 5.6.b.</u> The validity research should apply to the population for which the test is intended. It is known that validity coefficients can decrease considerably when a homogeneous group is used instead of a heterogeneous group (with respect to the variable measured). For example, it would be incorrect to validate a test intended for therapy selection on a sample of the general population. The sample has to be described by means of relevant psychological and demographic variables to facilitate a sound assessment of this item.

<u>Item 5.6.c.</u> Sometimes the choice of a criterion is obvious and criterion scores can easily be collected, while in other cases criterion measures have to be specially developed and collected. In all cases the criterion has to be described fully, and it should be indicated which relevant behavioural aspects are included in the criterion measure and which are not. This is especially true for composite criteria. When the intercorrelations of the separate components of a criterion are low, it is better to determine separate validity coefficients for each of the components than for the overall criterion only.

<u>Item 5.6.d.</u> Some examples of information which should be available when assessing the quality of the validation study are listed below.

- When the validity coefficient is corrected for attenuation or restriction of range, the uncorrected coefficient must be given as well, since these corrections often result in the overestimation or underestimation of the validity coefficient in specific cases. When the test has passed the developmental stage the formula for the correction of attenuation should in no way continue to be applied when correcting for unreliability of the test itself.
- Are the results of cross-validation provided?
- The size of the samples should be mentioned. The smaller the sample sizes, the wider the confidence intervals of the regression weights and the validity coefficients (and the more cross-validation is needed).

References

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, pp. 508-600). Washington, DC: American Council on Education.

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

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.

De Zeeuw, J. (1978). Algemene psychodiagnostiek II. Testtheorie [General psychodiagnostics II. Test theory]. Amsterdam, The Netherlands: Swets & Zeitlinger.

Drenth, P. J. D., & Sijtsma, K. (1990). Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen [Test theory. Introduction in the theory of the psychological test and it's applications]. Houten/Antwerpen, The Netherlands: Bohn Stafleu Van Loghum.

Evers, A. (2001). Improving Test Quality in the Netherlands: Results of 18 Years of Test Ratings. International Journal of Testing, 1(1).

Evers, A., Caminada, H., Koning, R., Ter Laak, J. , Van der Maesen de Sombreff, P., & Starren, J. (1988). Richtlijnen voor ontwikkeling en gebruik van psychologische tests en studietoetsen [Standards for the development and use of psychological and educational tests]. Amsterdam, The Netherlands: NIP.

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

Frisbie, D. A. (1988). Reliability of scores from teacher-made tests. Educational Measurement: Issues and Practice, 7, 25-35.

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

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

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. European Journal of Psychological Assessment, 10, 229-244.

Heuvelmans, A. P. J. M., & Sanders, P. F. (1993). Beoordelaarsovereenstemming [Interrater agreement]. In T. J. H. M. Eggen & P. F. Sanders (Eds.), Psychometrie in de praktijk [Psychometrics in practice](pp. 443-470).Arnhem, The Netherlands: CITO.

Hofstee, W. K. B., Campbell, W. H., Eppink, A., Evers, A., Joe, R. C., Van de Koppel, J. M. H., Zweers, H., Choenni, C. E. S., & Van der Zwan, T. J. (1990). Toepasbaarheid van psychologische tests bij allochtonen. LBR-reeks nr.11 [Applicability of psychological tests for ethnic minorities. LBR-series nr.11]. Utrecht, The Netherlands: LBR.

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, The Netherlands: Wolters-Noordhoff.

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

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.

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

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

Veldhuijzen, N. H., Goldebeld, P., & Sanders, P. F. Klassieke testtheorie en generaliseerbaarheidstheorie[Classical test theory and generalizability theory]. In T. J. H. M. Eggen & P. F. Sanders (Eds.), Psychometrie in de praktijk [Psychometrics in practice] (pp.33-82). Arnhem, The Netherlands: CITO.

Verstralen, H. H. F. M. (1993). Schalen, normen en cijfers [Scales, norms and grades]. In T. J. H. M. Eggen & P. F. Sanders (Eds.), Psychometrie in de praktijk [Psychometrics in practice] (pp. 471-509). Arnhem, The Netherlands: CITO.

Zegers, F. E. (1989). Het meten van overeenstemming [The measurement of agreement]. Nederlands Tijdschrift voor de Pyschologie, 44, 145-156.

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

Table 1

Rules for determining the final rating for Criterion 1 (Theoretical basis).

RULES	FINAL ASSESSMENT
· If the key question is rated ‘+’ and:	
- both items 1.2 and 1.3 are rated ‘+/-’ or ‘+’	GOOD
- either item 1.2 or item 1.3 is rated ‘-’	SUFFICIENT
- both items 1.2 and 1.3 are rated ‘-’	INSUFFICIENT
· If the key question is rated ‘+/-’ and:	
- both items 1.2 and 1.3 are rated ‘+/-’ or ‘+’	SUFFICIENT
- item 1.2 and/or item 1.3 are/is rated ‘-’	INSUFFICIENT
· If the key question is rated ‘-’:	INSUFFICIENT

Table 2A

Rules for determining the final rating for Criterion 2A (Test materials).

RULES	FINAL ASSESSMENT
· If all key questions are rated '+' and:	
- the sum of items 2.4, 2.5 and 2.6 > 1	GOOD
- the sum of items 2.4, 2.5 and 2.6 = 0, +0.5, or +1	SUFFICIENT
- the sum of items 2.4, 2.5 and 2.6 < 0	INSUFFICIENT
· If one of the key questions is rated '+/-' and:	
- the sum of items 2.4, 2.5 and 2.6 > 1	SUFFICIENT
- the sum of items 2.4, 2.5 and 2.6 ≤ 1	INSUFFICIENT
· If one of the key questions is rated '-':	INSUFFICIENT

Note. A '+'-rating is scored +1, a '±'-rating is scored 0, and a '–'-rating is scored –1. The scores for questions 2.4.a and 2.4.b have to be averaged before summing the scores.

Table 2B

Rules for determining the final rating for Chapter 2B (Test Manual).

RULES	FINAL ASSESSMENT
· If the key question is rated '+' and:	
- at least three of items 2.8 to 2.13 are rated '+'	GOOD
- less than three of items 2.8 to 2.13 are rated '+', and less than three of items 2.8 to 2.13 are rated '-'	SUFFICIENT
- at least three of items 2.8 to 2.13 are rated '-'	INSUFFICIENT
· If the key question is rated '-':	INSUFFICIENT

Table 3

Rules for determining the final rating for Criterion 3 (Norms).

RULES	FINAL ASSESSMENT
· If both key questions are rated ‘+’ and:	
- the sum of the other items ≥ 1	GOOD
- the sum of the other items ≤ 0	SUFFICIENT
· If key question 3.1 is rated ‘+’ and key question 3.2 is rated ‘+/-’ and:	
- the sum of the other items ≥ 1	SUFFICIENT
- the sum of the other items ≤ 0	INSUFFICIENT
· If one of the key questions is rated ‘-’	INSUFFICIENT

Note. A ‘+’-rating is scored +1, a ‘±’-rating is scored 0, and a ‘–’-rating is scored –1.

Table 4

Rules for determining the final rating for Criterion 4 (Reliability).

- If the key question is rated '-', the final rating will be INSUFFICIENT.
- If the key question is rated '+', the extent of the reliability is rated with item 4.2. This rating is a provisional assessment. The quality of the research design and the completeness of the information supplied, as rated in item 4.3, can give reason to adjust this rating downwards. For the final reliability rating, the coefficient which comes closest to the purpose of the test is of overriding importance when the results of various kinds of coefficients differ. For example, if the purpose is to predict over time, then an index of stability is more relevant than an internal consistency coefficient.

Table 5

Rules for determining the final ratings for Criteria 5A (Construct validity) and 5B (Criterion validity).

- If the key question for the respective criterion (i.e., item 5.1 for Criterion 5A, and item 5.4 for Criterion 5B) is rated ‘-’, the final rating is INSUFFICIENT.
- If the key question is rated ‘+’, the quality of the evidence provided is rated with item 5.2 (for construct validity) and 5.5 (for criterion validity). This rating is a provisional assessment. The quality of the research design and the completeness of the information supplied, as rated in items 5.3 and/or 5.6, can give reason to adjust these ratings downwards.

Author Notes

Correspondence concerning this article should be addressed to Prof. Arne Evers, University of Amsterdam, Department of Psychology, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands, or by email to <evers@psy.uva.nl>. Due to space restrictions, the format used in this article differs from the format of the original Dutch version (Evers, Van Vliet-Mulder, & Groot, 2000).

Acknowledgements

The final text of this revision of the Dutch rating system for test quality was formulated for COTAN by Arne Evers. The author is grateful for the contributions and feedback of the following COTAN-members who participated in the process of revision:
Drs. A.E. Akkerman, Drs. R.J. van Alphen de Veer, Prof. dr. N. Bleichrodt, Dr. J.B. Blok, Dr. H. van Boxtel, Dr. M.H.M. Breteler, Drs. M.E. Dinger, Drs. M.I.A. Douma, Drs. A. Gazendam, Drs. C.J. Groot, Drs. E.F.M. Pouw, Dr. W.C.M. Resing, Prof. dr. P.F. Sanders, Drs. A.M.H.L. Serlier-van den Bergh, Dr. N. Smid, Dr. J.C.M.G. Starren, J.C. van Vliet-Mulder, Prof. dr. W. Wijnen and Dr. J.N. Zaal. From outside COTAN dr. P.J. Tellegen provided very valuable input.