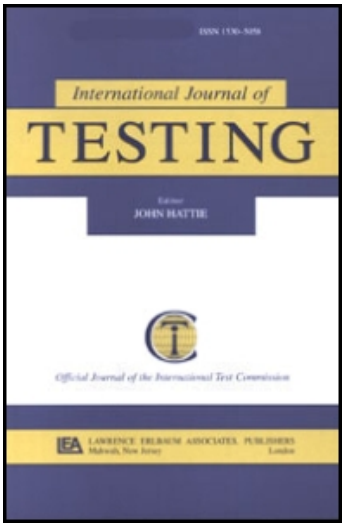

International Journal of Testing
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653658

The Dutch Review Process for Evaluating the Quality of Psychological Tests: History, Procedure, and Results

Arne Evers[a]; Klaas Sijtsma[b]; Wouter Lucassen[c]; Rob R. Meijer[d]
[a] University of Amsterdam, [b] Tilburg University, [c] Meurs HRM, [d] University of Groningen,

Online publication date: 06 November 2010

To cite this Article Evers, Arne , Sijtsma, Klaas , Lucassen, Wouter and Meijer, Rob R.(2010) 'The Dutch Review Process for Evaluating the Quality of Psychological Tests: History, Procedure, and Results', International Journal of Testing, 10: 4, 295 — 317
To link to this Article: DOI: 10.1080/15305058.2010.518325
URL: http://dx.doi.org/10.1080/15305058.2010.518325

International Journal of Testing, 10: 295–317, 2010

ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305058.2010.518325

Routledge
Taylor & Francis Group

The Dutch Review Process for Evaluating the Quality of Psychological Tests: History, Procedure, and Results

Arne Evers
University of Amsterdam

Klaas Sijtsma
Tilburg University

Wouter Lucassen
Meurs HRM

Rob R. Meijer
University of Groningen

This article describes the 2009 revision of the Dutch Rating System for Test Quality and presents the results of test ratings from almost 30 years. The rating system evaluates the quality of a test on seven criteria: theoretical basis, quality of the testing materials, comprehensiveness of the manual, norms, reliability, construct validity, and criterion validity. The update concentrated on two main issues. First, the texts of all criteria were adapted to apply to both paper-and-pencil tests and computer-based tests. Second, the items and the recommendations with respect to the seven criteria were extended to include new developments. The most important extensions are item response theory for test development, continuous norming, domain-referenced interpretation and criterion-referenced interpretation, and the use of non-traditional, modern types of reliability estimation. Longitudinal results show a steady increase of average test quality, but the quality of the norms and the (lack of) research on criterion validity still appear to be a matter of concern.

Keywords: Dutch Committee on Testing, test documentation, test quality, test review system

All authors are members of the Dutch Committee on Testing (COTAN).

Correspondence should be sent to Arne Evers, Work & Organizational Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands. E-mail: a.v.a.m.evers@uva.nl

Quality assessment of products and services is very important in current society. A Google search using the keyword "quality" results in tens of thousands of hits revealing a great variety of topics, from the quality of colleges and nursing homes to that of parking garages and the water in ponds. When consulting the website of the International Organisation for Standardisation (www.iso.org), the Dutch Normalization Institute (www.nen.nl), and websites of comparable organizations in other countries, one finds that quality criteria have been developed for thousands of products and services. Furthermore, consumer organizations in many countries seeking to protect people from corporate abuse exercise the policy of publishing tests of product quality in their monthly journals.

Concern about quality assessment of many products only has a short history. For example, a recent initiative related to the field of testing is the development of an ISO standard for assessment procedures in work settings (Born, 2009). However, concern about the quality of tests and accompanying materials dates back to 1895, when the first Committee on Mental Measurements of the American Psychological Association was installed. In 1938, the first edition of a still continuing series of Buros' Mental Measurements Yearbooks was published (Buros, 1938; Spies, Carlson, & Geisinger, 2010). The Buros' Yearbooks contain reviews of commercially available educational and psychological tests in the English language, covering a period of more than 70 years. In this period thousands of comprehensive and rather *holistic* test reviews were published.

Forty years ago, the Dutch Committee on Testing (COTAN) of the Dutch Psychological Association (NIP) was the first to publish a book containing *ratings* of the quality of psychological tests (NIP, 1969). Five updates have followed since, most recently in 2009 (Evers, Braak, Frima, & Van Vliet-Mulder, 2009). Surprisingly, it took a long time before other countries dealt with the evaluation of test quality in a systematic way. In the United Kingdom, test evaluations were published regularly since 1990 (Bartram, Lindley, & Foster, 1990). Up to 2009, 70 tests have been published and registered (Lindley, 2009). Norway followed in 2005 (Nielsen, 2009; 8 evaluated tests) and Germany in 2006 (Moosbrugger et al., 2009; 13 evaluated tests). In both countries, tests are assessed nowadays using a fully operational system for the evaluation of their quality. Brazil (Wechsler, 2007), Spain (Prieto & Muñiz, 2000), and Sweden (Tideman, 2007) have implemented or are preparing the initiation of such a system, but no results are known yet. In the United Kingdom, Norway, Spain, and Sweden, a European model for assessing test quality is or will be used (http://www.efpa.eu/professional-development/tests-and-testing). The European model is inspired by previously existing models, such as the British (Bartram, 1996; Bartram et al., 1990), the Dutch (Evers, 2001a, 2001b), and the Spanish models (Prieto & Muñiz, 2000).

Considering the importance of psychological tests for the work of psychologists and the impact tests may have on their clients, this overview is a bit disappointing,

both with respect to the number of countries that practice quality control and the number of evaluated tests. For example, a survey on test use in The Netherlands (Evers, Zaal, & Evers, 2002) revealed that more than 800 different tests were used, and there is no reason to assume that this number is considerably smaller in other countries. We conclude that only a small portion of the tests that are used in practice has been evaluated thus far. In The Netherlands, with its longer history of test evaluations, between 1982 and 2010, 878 tests have been evaluated, including double counts concerning tests that were evaluated twice (e.g., the Wechsler Adult Intelligence Scale [WAIS] and the WAIS-III, and the Minnesota Multiphasic Personality Inventory [MMPI] and the MMPI-2).

A description of the previous Dutch rating system for test quality was published in *International Journal of Testing* (Evers, 2001a, 2001b). The main objective of the present paper is to describe the major revision of the system, which was completed in 2009 (Evers, Lucassen, Meijer, & Sijtsma, 2009). In addition, we present data with respect to the quality of the Dutch test repertory and the results of almost 30 years of test ratings.

THE DUTCH RATING SYSTEM: HISTORY AND DESCRIPTION

As a result of the growing awareness of the importance "to promote the better use of better tests," the NIP founded the COTAN in 1959. Representatives of each of the nine departments of psychology of Dutch universities and representatives of psychologists practicing in various areas were appointed as members. Today, test constructors, test users, publishers, and educational and government organizations consider the COTAN an independent authority in the field of testing. In The Netherlands, no specific legislation concerning the use of tests—such as laws concerning who is entitled to use tests—or the quality of tests exists. Therefore, the COTAN adopted the information approach, which entails the policy of improving test use and test quality by informing test constructors, test users, and test publishers about the availability, the content, and the quality of tests. The two key instruments used are the *Documentation of Tests and Test Research* and the system for rating test quality.

The *Documentation of Tests and Test Research* contains a description of all tests available in The Netherlands and contains excerpts of research conducted with these tests. A criterion for being listed in the book is that at least some research in Dutch samples on norms, reliability, or validity has been published. The first edition was published in 1961 and contained descriptions of 247 tests. New editions appeared with intervals of 4 to 10 years, in the meantime supplemented by loose-leaf supplements. The most recent edition (Evers, Braak, et al., 2009) is digitalized

and is updated monthly, containing descriptions and quality assessments of more than 650 tests. The descriptions have a fixed format and concern eight characteristics: target group(s), construct(s) to be measured, availability/publisher, administration procedure, time needed for administration, content (e.g., number and type of items), scoring (way of scoring, available norm groups, type of norms), and references.

In 1969 (NIP, 1969), a major improvement in the test descriptions was the introduction of quality ratings for each test. These ratings referred to the overall quality of a test, which was summarized by ratings running from A (excellent) through F (inadequate test construction or theory underlying the test). For example, an A was given to a test with "an objective scoring system *and* norms on defined groups *and* positive results on reliability *and* sufficient results on construct and predictive validity" (NIP, p. 9, emphasis added).

Another big step forward was the introduction of a completely new system in 1980. This system was based on the *Standards for Educational and Psychological Tests* (American Psychological Association, 1974). The main reason for the development of this new system was dissatisfaction with the possibility that in the old system the same rating could be based on different kinds of input (e.g., either sufficient reliabilities or sufficient data on construct validity). Another reason was that the original system ignored interesting and relevant practical aspects, such as quality of scoring keys, instruction, and case descriptions. The new system assessed the quality of a test on five criteria: the theoretical basis of the test, the quality of the test materials and the manual, norms, reliability, and validity. The rating on each criterion could be "insufficient," "sufficient," or "good." A detailed questionnaire was developed that contained pre-coded answers, recommendations, and a weighing system to determine the final rating for each criterion, and all tests that had already been rated were re-rated using this system.

In 1997, a revised version of this system was put into use and was published in *International Journal of Testing* (Evers, 2001b). The revision contained an update of quality requirements for items and recommendations, but the main change was the extension of the five criteria to seven criteria in an effort to further enhance the information and communication value of the system. The criteria "quality of the test materials and the test manual" and "validity" were divided into "quality of the test materials" and "comprehensiveness of the manual," and "construct validity" and "criterion validity," respectively. In the 1997 revision, relatively little attention was paid to new developments with respect to test construction, such as item response theory (IRT), and test administration, such as computer-based testing. The reason was that although these developments were already well known at that time, their application in practical test construction was almost absent in The Netherlands. Hence, there seemed to be no need to adapt the rating system to these developments. However, this situation has rapidly changed during the past decade, urging a revision with respect to content.

For a full understanding of the COTAN review system, first attention is given to some characteristics that have not changed. Next, the main changes that were implemented in 2009 are described.

WHAT DID NOT CHANGE?

The general structure and the assessment criteria were left unchanged. As in the previous version, each criterion is assessed using several items, including one or more key questions. A negative answer to a key question directly leads to the rating "insufficient" for the criterion involved. To support answering the items, extensive guidelines are given to judges. In addition, for each criterion directions for combining item scores into a total score are supplied, and this total score serves to decide on the rating for a criterion, which can be insufficient, sufficient, or good.

The rating procedure was also left unchanged. A distinctive feature of this procedure is that two anonymous, independent raters evaluate each test, which is the usual procedure for the review of manuscripts for journals. Raters are experts with respect to the specific type of test, the theory underlying test construction, and the practical use of the test. Raters never review their own tests, tests of colleagues in the same organization, or tests involving other circumstances, which may cause a conflict of interest. When discrepancies in the final ratings on any of the seven criteria show up, the raters have to discuss their ratings and reach agreement. The senior editor of the COTAN monitors this process, if necessary asks a third rater, integrates the comments of the raters, and makes the final decision. Then, the ratings and comments are sent to the author who is given the opportunity to react to the outcome. In some cases (fewer than 5%) this may lead to an adjustment. Finally, the rating is published in the *Documentation*.

WHAT'S NEW IN THE 2009 REVISION?

In the first and later versions of the rating system, the questions and recommendations regarding the psychometric criteria (norms, reliability, and validity) were essentially based on classical test theory. The focus of the more practical criteria (quality of the test materials and the manual) was primarily on paper-and-pencil tests. For a long time, these choices did not limit the applicability of the system. Some rare tests were constructed using non-classical approaches, but assessment could take place if raters were prepared to improvise, and they were. However, when more tests appeared that were constructed using non-classical approaches, this became problematic because test constructors did not know well which information they had to supply for the assessment of their tests.

The update of the system concentrated on two main issues. First, the texts of all criteria were adapted to apply to both paper-and-pencil tests and computer-based

tests. Second, for each criterion the items and recommendations were extended to include new developments (or revived old techniques). The most important are item response theory for test development, continuous norming, domain-referenced interpretation and criterion-referenced interpretation, and the use of other types of reliability estimation methods than the traditional methods using the test-retest correlation and the alpha coefficient. Keuning (2004) provided much of the input for the improvements that deal with computer-based tests. His work was based, among others, on the *Guidelines on Computer-based and Internet Delivered Testing* (International Test Commission, 2003), and Parshall, Spray, Kalohn, and Davey (2002). Wools, Sanders, and Roelofs (2007) developed a system for the evaluation of instruments for competence assessment, which provided the input for the improvements concerning domain-referenced interpretation and criterion-referenced interpretation.

The texts of the items for each criterion are included in the Appendix. Due to limited space, the recommendations and directions for determining the final ratings are not included.[1] Next, we describe the main ideas underlying the assessment criteria, the items, and the recommendations used to help raters. We provide special attention to the implemented changes.

DESCRIPTION OF EVALUATION CRITERIA

Theoretical Basis of the Test

The information provided for this criterion should enable the prospective test user to judge whether the test is suitable for his or her purposes. It contains three items that deal with the theoretical basis and the logic of the test development procedure. The key item (1.1)[2] asks whether the test manual clarifies the construct that the test purports to measure, the groups for which the test is meant, and the application of the test. The answer to this item determines the demands on the research published: the more ambitious the test author's claims, the greater his or her obligation to deliver empirical norms and evidence of validity. The second item (1.2) deals with the theoretical elaboration of the test construction process: Is the test based on an existing psychological theory or on new ideas that have brought about changes in this theory? Assessment of this item also involves translated tests or adaptations of a foreign instrument. In addition, the manual has to supply definitions of the constructs to be measured. If both a paper-and-pencil version and a computer-based test version exist (or if the computer-based test version is an adaptation of the paper-and-pencil version), differences in item content or item wording must be specified. The third item (1.3) asks for information about the operationalization of the construct. This item also deals with the content validity of the test, in particular for educational tests. Because in adaptive tests different persons are administered different sets of items, it is possible that specific subjects

are under- or overrepresented. Therefore, for adaptive tests it is required to specify how the representativeness of the test content is guaranteed (e.g., see Kingsbury & Zara, 1991).

Quality of the Test Materials

The new rating system contains separate sets of items for paper-and-pencil tests and computer-based tests. If a test can be administered both ways, both sets of items have to be answered (and two ratings for this criterion are published). Both sets contain eight items, of which three are key items. The key items deal with the standardization of the test content (2.1), the objectivity of the scoring system (2.2), and the presence of unnecessary culture-bound words or content that may be offensive to specific ethnic, gender, or other groups (2.3). The other five items refer to the design, the content and the form of the test materials (2.4 and 2.7), the instructions for the test taker (2.5), the quality of the items (2.6), and the scoring of the items (2.8).

In adaptive tests, test takers receive different sets of items. Standardization that assumes the same test for each testee does not apply to adaptive tests. Instead, for adaptive tests it is required that the decision rules for the selection of the next item are specified (2.9). For tests that are computer scored, information has to be provided that enables the rater to check the correctness of the scoring (2.10). Other deviations relative to the items with respect to paper-and-pencil administration are that, for computer-based tests, special attention is given to the resistance of the software to user errors (2.12), the quality of the design of the user interface (2.15), and the security of the test materials and the test results (2.16).

Comprehensiveness of the Manual

This criterion evaluates the comprehensiveness of the information the manual provides to the test user to enable the well-founded and responsible use of the test. The starting point is that the manual, in addition to all the information that the test user needs to administer and interpret the test (commonly this is called the User's Guide), should also supply a summary of the construction process and the relevant research (commonly called the Technical Manual). This information should enable the test user and the test reviewer to form an opinion about the psychometric quality of the test.

This criterion contains seven items, of which one is a key item. The key item asks whether there is a manual at all (3.1). The other items deal with the completeness of the instructions for successful test administration (3.2), information on restrictions for the use of the test (3.3), the availability of a summary of results of research performed with the test (3.4), the inclusion of case descriptions (3.5),

the availability of indications for test-score interpretation (3.6), and statements on user qualifications (3.7) .

For computer-based tests three extra items are available. The first item asks whether sufficient information is supplied with respect to the installation of the software (3.8), the second item asks whether there is sufficient information regarding the operation of the software and the opportunities provided by the software (3.9), and the third item asks for the availability of technical support for practical software use (3.10).

Norms

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

The criterion Norms is assessed using two key items and three separate sections on norm-referenced, domain-referenced, and criterion-referenced interpretation. The two key items apply to all sections. The first item (4.1) checks whether norms, standards, or cut scores are provided. The second item (4.2) asks in which year or period the data were collected. An important innovation is the addition of the comment "the norms are out-dated" when norms are older than 15 years. This warning should alert test users to be careful interpreting the norms. When the norms are older than 20 years the comment, "because the norms are out of date they are not useful anymore," is added in combination with a downgrading of the rating to "insufficient." For this reason, once per year the norms of all tests are re-evaluated. If test authors have the opinion that the norms for their test have not changed after 15 or 20 years, they have to document this claim by providing

research data, for example by showing that mean scores for a specific subgroup did not change. Although in the formulation of item 4.2 only the word "norms" is used, this item applies also to standards and cut scores.

Five items refer to norm-referenced interpretation, one of which is a key item. This item (4.3) deals with the size and the representativeness of the norm group or norm groups. Since the first version of the rating system, clear-cut rules were formulated with respect to the size of the norm groups. For tests intended for making important decisions,[3] a norm group smaller than 300 is considered "insufficient," between 300 and 400 "sufficient," and larger than 400 "good." For tests intended for making less important decisions, corresponding group sizes are 200 and 300, respectively. These rules apply to the "classical norming" approach in which norms are constructed for separate (age or year) groups. The continuous-norming procedure uses the information from all available groups to construct the norms for a specific group, which results in more accurate norms than classical norms (e.g., Zachary & Gorsuch, 1985). Thus, a continuous-norming procedure produces the same accuracy using smaller individual norm groups.

Bechger, Hemker, and Maris (2009) studied a continuous-norming approach for eight groups and developed rules for the size of individual norm groups to be used in continuous norming. They used a linear regression approach assuming equal variances and standard-normal score distributions in all groups. To compare the accuracy of both approaches, they used the standard error of the mean. The results showed that a group size of about 70 in the continuous approach (for eight groups) produced the same accuracy as a group size of 200 in the classical approach, and that group sizes of 100 and 150 corresponded to sizes of 300 and 400, respectively. These group sizes are mean values, but in the outer groups accuracy is a bit worse than in the middle groups; hence, the outer groups should be larger whereas the middle groups may be smaller. The computation of these group sizes is difficult for various numbers of groups and for all existing continuous-norming approaches (although linear regression is the most common approach), and test constructors are advised to supply evidence about the level of accuracy of the continuous norms for their test. They should also supply information on the other moments of the score distribution as well as information about deviations from the statistical assumptions underlying their continuous-norming procedure.

Concerning the representativeness of the norm groups, test constructors should at least provide evidence with respect to age, gender, ethnic group, and region. Convenience samples must be described with respect to variables that can be considered to be relevant, such as educational level and type of job for norms in personnel selection tests. The other four items for this criterion ask for information on the norm scale used (4.4); means, standard deviations, and other information with respect to the score distributions (4.5); differences between various subgroups (4.6); and the standard error of measurement, the standard error of estimate, or the test information function (4.7).

Three items refer to domain-referenced interpretation and criterion-referenced interpretation. In domain-referenced interpretation, different standard-setting methods can be used (e.g., Cascio & Aquinis, 2005; Hambleton, Jaeger, Plake, & Mills, 2000; Vos & Knuver, 2000). The specific method chosen, the procedures for determining the cut score(s) (4.9), and the training and selection procedure of the judges (4.10) have to be described. Significant importance is assigned to inter-rater agreement with respect to the determination of the critical score (4.8). Following Shrout (1998), kappa coefficients and intra-class correlation coefficients below .60 are assessed to be "insufficient," values between .60 and .80 "sufficient," and values above .80 "good."

In criterion-referenced interpretation, cut scores or expectancy tables are derived from empirical research. Actually, this concerns research on the criterion validity of the test, which serves setting norms empirically. Examples are research on predictive validity of a test in personnel selection and research on the sensitivity and specificity of a test in clinical psychology. This type of research is evaluated here exclusively from the perspective of setting norms; criterion validity per se is evaluated in the section concerned. The three items check whether the results show sufficient validity (4.11), whether the sample used is comparable to the population in which the test is used (4.12), and whether the sample is sufficiently large (more than 200 respondents is "sufficient," more than 300 respondents is "good") (4.13).

Reliability

Reliability is a basic requirement for a test. However, different estimation methods may produce different reliability estimates, and in different groups the test score may have different reliabilities. Thus, reliability results should be evaluated from the perspective of the test's application. Classical test theory assumes that a test score additively consists of a reliable component (also called true score) and a component caused by random measurement error. The objective of the reliability analysis is to estimate the degree to which test-score variance is due to true-score variance. In the previous version of the rating system, recommendations were given with respect to parallel-form reliability, reliability on the basis of inter-item covariances, test-retest reliability, and inter-rater reliability. In the new version, methods from IRT, generalizability theory, and structural equation models are also mentioned. Although reliability estimates can differ depending on method (particularly important is the composition of the error components) and the characteristics of the group studied (particularly influential is the test-score variance), for the reliability criterion only one qualification ("insufficient," "sufficient," or "good") is given. This qualification is based on the range (described later) of the majority of the coefficients. A footnote can be added to this qualification to mention exceptions, for example "the reliability of Scale X is insufficient" or "the reliability in Group Y is insufficient."

The reliability criterion has three items, of which one is a key item. The key item (5.1) checks whether any reliability results are provided at all. The second question (5.2) asks for the level of the reliability. For tests intended for making important decisions, reliability lower than .80 is considered "insufficient," between .80 and .90 "sufficient," and above .90 "good." For tests intended for making less important decisions, corresponding boundary values are .70 and .80, respectively. For reliability estimates based on IRT, generalizability theory, or structural equation modelling, comparable values are required. For IRT, methods for the reliability of the estimated latent ability or trait (Embretson & Reise, 2000) or the unweighed total score (*rho*, Mokken, 1971) are available. Alternatively, the test information function may be used (e.g., Reise & Havilund, 2005).

An important change concerns reliability estimation on the basis of inter-item covariances. It is almost standard that test constructors report coefficient alpha (Cronbach, 1951). However, alpha underestimates the reliability of a test (e.g., Novick & Lewis, 1967), and other methods such as Guttman's lambda2 (Guttman, 1945) and the greatest lower bound (*glb*; Ten Berge & Sočan, 2004) give higher underestimates that come closer to the real reliability of the test (Sijtsma, 2009a, 2009b)[4]. In the recommendations it is now explicitly allowed, and even encouraged, to use these and other alternative methods.

The third item (5.3) deals with the quality of the reliability research design and the completeness of the information supplied. The rating for reliability can be adjusted downwards when this research shows serious weaknesses. For example, reliability should be computed for the norm samples, and if that is not feasible, for groups that are comparable to the norm samples. When the groups used are more heterogeneous than the norm samples, generally the reliability coefficient is inflated.

Construct Validity

Validity is the extent to which a test fulfils its purpose (Drenth & Sijtsma, 2006). In the current validity conception, different forms of evidence on the validity of tests should not be considered to represent distinct types of validity, but validity should be considered a "unitary concept" (American Educational Research Association et al., 1999, p. 11). From this point of view, it is important to collect evidence of validity that supports the intended interpretation and proposed use of the test's scores (Ter Laak & De Goede, 2003). When the purpose is description, other validity information is required than whether the purpose is prediction or classification. However, for a standardized rating procedure it is necessary to structure the concept of validity and thereby the rating process. Therefore, in the revised version of the rating system the distinction between construct validity and criterion validity as separate criteria is maintained. Types of validity evidence that are construct-related (described later) are required for almost all tests, whatever

the purpose of test use (even when the purpose of a test is mere prediction, it would be odd not wanting to know what the test actually measures). Types of validity evidence that are criterion-related (see next section) will not be required for tests that are not intended for prediction.

Although the terminology used in this rating system differs from the 1999-Standards, this seems merely a difference in structuring the framework. In the 1999-Standards evidence for validity based on the internal structure and evidence based on relations to other variables (external structure) is distinguished. In the COTAN-system, all research with respect to the internal structure is considered relevant for construct validity. Research concerning the external structure may be relevant for either construct validity (e.g., convergent and discriminant evidence) or for criterion validity (primarily predictive studies dealing with test-criterion relationships). Sometimes predictive research can also add to the construct-related evidence (6.2.f).

Construct-related evidence should support the claim that the test measures the intended trait or ability. This concerns answers to questions such as "What does the test measure?" and "Does the test measure the intended concept or does it partly or mainly measure something else?" As a consequence of the diversity of validity research, in the recommendations of the former version of the rating system few directions were given with respect to the type and the comprehensiveness of research that would be enough for the qualification "sufficient" or "good." However, after 30 years of experience with the evaluation of research on construct validity it is much clearer now what kinds of research are usually performed. On the basis of this experience, the rating system distinguishes six types of research in support of construct validity. These types are: research on the dimensionality of the item scores (e.g., evidence from factor analysis of the data), the psychometric quality of the items (e.g., evidence from corrected item-total correlations or discrimination parameters), invariance of the factor structure and possible bias (e.g., evidence for fair test use), convergent and discriminant validity (i.e., supporting construct validity), differences between relevant groups (e.g., clinical and normal groups), and other research (e.g., research on criterion validity that is also relevant for construct validity).

For each type of research, the rating system gives extensive directions, for example, concerning the level of the corrected item-total correlations or the discrimination parameters, the standard error of the difficulty parameters of items in IRT-models, and the desired sample size. In addition, the system distinguishes research on the internal test structure and research on the external test structure. The first three types of research supply evidence on the internal structure, and the last three types on the external structure. Evidence in both categories is required to earn the qualification "sufficient" (or "good").

The structure of the items with respect to construct validity is the same as for reliability. First, the provision of results is ascertained by means of a key item (6.1).

Second, the sufficiency of the evidence supporting construct validity is assessed (6.2). Third, the quality of the research design is assessed (6.3). This may lead to a downward adjustment of the rating. For example, attention has to be paid to chance correlations in the absence of a priori hypotheses, the reliability of the convergent and discriminant measures, and the sample sizes.

Criterion Validity

Research on criterion-related evidence should demonstrate that a test score is a good predictor of non-test behavior or outcome criteria. Prediction can focus on the past (retrospective validity), the same moment in time (concurrent validity), or on the future (predictive validity). Basically, evidence of criterion validity is required for all kinds of tests. However, when it is explicitly stated in the manual that test use does not serve prediction purposes (such as educational tests that measure progress), criterion validity is not evaluated and instead a rating receives the comment "according to the author/publisher this test is not meant for predictive purposes; hence, criterion validity is not applicable."

The structure of the items with respect to criterion validity is the same as for construct validity. The recommendations (of item 7.2) contain a section on the use of ROC-curves (Receiver Operating Characteristic). Also, an overview of validity coefficients relevant in personnel selection (based on Schmidt & Hunter, 1998) and ROC-values (based on Swets, 1988) is supplied so as to guide the raters in formulating their judgment. The use of validity generalization is allowed, provided the author presents a reasonable case for the similarity of the situations for which the generalization is claimed. For translated tests, results of foreign studies may be generalized only if equivalence (e.g., ascertained by means of confirmatory factor analysis) of the original and the translated version has been shown. Deficiencies in the research design (e.g., inadequate research groups, too small sample sizes, inadequate criterion measures) may lead to downwards adjustment of the rating (7.3).

QUALITY OF DUTCH TESTS

Using the seven assessment criteria, in this section we discuss results with respect to the quality of Dutch tests in 2009. In addition, we provide the mean quality assessments of tests included in the *Documentation of Tests and Test Research* in 1982, 1992, 2000, and 2009. In the *Documentation*, the assessments of research instruments (i.e., tests to be used for research but not for individual diagnostics) and tests of Flemish-Belgian origin (these tests are also in Dutch) are also included. Research instruments usually are published in journals, and for nearly all of them standard test material, a manual, and norms are absent. Flemish-Belgian tests

are used in The Netherlands without translation or adaptation, but usually Dutch norms are absent. Both types of tests were excluded from our analyses here so that the numbers of tests used for these analyses (236, 299, 372, and 540 tests, respectively) are smaller than the numbers of tests included in the *Documentation* (278, 372, 457, and 622 tests, respectively). Including these test types would not give a fair picture of the quality of Dutch tests.

In the first four columns of Table 1, results on Dutch test quality in 2009 are presented. For five of the seven criteria, the quality of a considerable majority of the tests (about two thirds or more) is at least "sufficient." However, for two criteria (Norms and Criterion validity) the quality of the majority of tests is "insufficient." For 9 (2%) tests, norms are considered to be irrelevant, because in the manual only intra-individual comparison of test scores is recommended. For 67 (12%) tests, which are mainly educational, criterion validity is considered to be irrelevant because prediction is not envisaged (as in pupil monitoring and evaluation systems).

To compare the quality of the tests at the four reference dates, the mean quality of a test was computed by averaging the five (1982 and 1992) or seven (2000 and 2009) ratings. To compute the mean score for a test, on each criterion "insufficient" was given the value 1, "sufficient" the value 2, and "good" the value 3. Consequently, per year the mean of these means was computed (see Table 2). A steady increase in Dutch test quality can be observed, although this trend seems to be weakening.

The means in Table 2 show a general trend, which can be analyzed with the more detailed information in Table 1. The results show that the improvement of general test quality over the four reference years can be attributed to improvements on six criteria but not on criterion validity. Until 2000, there was a clear increase in quality on four criteria—theoretical basis, quality of testing materials, reliability and construct validity—but the improvement seems to level off after 2000. For comprehensiveness of the manual and norms, the quality ratings show a substantial improvement between 2000 and 2009.

Evers (2001a) performed some follow-up analyses to investigate whether the general improvement could be attributed to the inclusion of better tests and the removal of worse tests or to an improvement of existing tests between two reference dates. Both processes can cause the improvement because tests are removed from the *Documentation* when they fall into disuse or when they are not available anymore and tests are re-rated when a general revision has been done, norms have been updated, new research has been published, or a new manual has been released. In 2001, it was concluded that both processes affected the results but that the effect on mean quality of an update of the test repertory was much stronger than the effect of maintenance and revision of existing tests. Because the urge to remove tests in the digital 2009-edition of the *Documentation* was not as strong as in the older book editions, only a few tests were removed. Hence, in 2009 the

TABLE 1
Test Quality in 2009, 2000, 1992, and 1982 on Seven (2009 and 2000) or Five (1992 and 1982) Assessment Criteria
Entries are Percentages of Documented Tests

	2009 ($N = 540$)			2000 ($N = 372$)			1992 ($N = 299$)			1982 ($N = 236$)		
Criteria	good	suff.	insuff.	good	suff.	insuff.	good	suff.	insuff.	good	suff.	insuff.
Theoretical basis	62	25	12	66	21	12	59	22	18	53	25	22
Quality of testing materials	70	22	8	70	22	8	55	32	13	47	35	18
Comprehensiveness manual	51	26	23	45	27	28						
Norms[1]	14	28	56	13	26	59	13	20	65	10	19	69
Reliability	27	42	32	27	41	33	21	40	39	21	33	46
Construct validity	18	48	34	18	47	35	9	41	50	8	35	57
Criterion validity	7	21	59[2]	8	25	67						

[1]For 1% to 2% of the tests the criterion 'Norms' was not applicable.
[2]For 12% of the tests Criterion validity' was not applicable.

TABLE 2
Mean Test Quality From 1982 until 2009

Year	Mean Test Quality*	Number of Tests
1982	1.84	236
1992	1.94	299
2000	2.01	372
2009	2.03	540

**Note.* Insufficient = 1, sufficient = 2, good = 3.

effect of deleting old tests on mean quality can be ignored. This may explain at least partly the slowing down of quality improvement.

DISCUSSION

Progressively gained insight in test theory, test construction, and test development has inspired us to change the content of the seven assessment criteria of the Dutch system for evaluating test quality. Consequently, the rules to determine the seven ratings also have been changed. This change has resulted in a system for assessing test quality that is up-to-date. An important collateral effect is that by translating these new developments into items and recommendations, the assignment for the reviewer is more structured and easier to handle. For the test author, the system is more transparent and it is clearer which information has to be supplied. A drawback of this thorough revision may be that the ratings produced by the previous system are not equivalent to those produced by the update.

Many changes are explicit formulations of informal rating practices already longer in use, and new elements were also introduced. The greater degree of specificity may alert raters and change assessment practices. A mini-experiment was done for the first five tests that were available for rating after the revised rating system was finished. For each test, two raters used the previous system and two other raters independently used the updated system. Except for two criteria, the ratings did not show systematic differences. For theoretical basis, the more-detailed information requested in the key item may in some cases lead to a more severe judgment. For norms, in continuous norming procedures the specification of rules for minimum numbers of respondents led to lower ratings when applied to tests initially evaluated by means of the previous system. This concerns a minority of all tests evaluated, but we conclude that ratings for norms based on the updated system cannot be compared to ratings based on the previous system. From today's perspective, older ratings of norms can best be considered a reflection of norms quality given knowledge of test theory at the time of evaluation.

Our comparison of the results over the years replicates Evers' (2001a) conclusion. The overall picture is positive showing that the quality of the test repertory is gradually improving. The large percentages of insufficient assessments for norms and criterion validity are a point of concern. For norms, quality is improving but progress is slow and the number of tests with low quality norms is still too high. The insufficient assessments for norms are often due to insufficient sample size, non-representativeness of the sample, or incomplete description of the sample, the population and/or the sampling procedure implying that representativeness cannot be determined. The first two shortcomings may be due to a limited financial budget and may be difficult for test constructors to overcome. However, one could argue, and this is the COTAN's point of view in this matter, that unreliable norms may render any investment in a test useless. Reliable norms are the basis for good decisions, and researchers are advised to invest their resources so as to optimize the quality of their norms. A positive development is that the assessment results show improvement with respect to description of the sample or the sampling procedure.

For criterion validity, the great number of insufficient assessments is due to the absence of any research whatsoever. Maybe testing agencies consider conducting such studies too demanding. As with norms research, this kind of research is expensive and hampered by many methodological problems (e.g., different types of restriction of range, which are most difficult to control). Cooperation by multiple test constructors and/or publishers in collecting these kinds of data may provide a solution for this problem. A Dutch initiative of the Taskforce Advising on Instruments for Indication in Special Education shows that this approach can be successful (Resing et al., 2008).

The publication of test ratings is an effective way of providing information on test quality to test users, test authors, and others involved in the field of testing. Every week, COTAN receives multiple requests for information on the quality of specific tests. Test constructors seek to fulfill the requirements of the rating system. Some governmental institutions require agencies they sponsor to use only tests the COTAN qualifies at least sufficient, and even in commercial tenders companies often require "COTAN-proof" tests. Most Dutch psychology departments use the COTAN rating system in their courses on test use and diagnostics.

As the COTAN has no interest in the outcome of a test assessment—for example, there is no formal relation with test publishers—independence of the assessment is guaranteed. The COTAN's independence is a prerequisite for acceptance of test ratings. The COTAN's continued efforts may have contributed to the gradual improvement of test quality in The Netherlands. Other influential factors are improved research facilities (e.g., online data collection, more computing power, easily accessible statistical software), improved knowledge of test construction methodology, and the activities of international organizations such as the International Test Commission and the Standing Committee on Tests and Testing of the European Federation of Psychologists' Associations.

ACKNOWLEDGEMENTS

The authors thank the COTAN-members R. H. van den Berg, J. B. Blok, M. Ph. Born, H. W. van Boxtel, M. Braak, M. E. Dinger, R. M. Frima, B. T. Hemker, P. P. M. Hurks, W. W. Kersten, E. F. M. Pouw, W. C. M. Resing, J. D. L. M. Schutijser, and T. van Strien for their valuable comments and participation in the process of revision. In addition, T. Bechger, J. Keuning, G. Maris, E. Roelofs, P. F. Sanders, and S. Wools provided valuable input.

NOTES

1. The complete system (at the moment only available in Dutch) can be downloaded from www.cotan.nl.
2. Throughout this text the numbers in brackets refer to the item numbers in the Appendix.
3. Important decisions are defined as decisions that are essential or irreversible and on which the test taker has little influence, such as in personnel selection or in placement in special educational programs.
4. Reliability coefficients that are not based on inter-item relations, such as a coefficient based on a test-retest procedure, can be lower than alpha.

REFERENCES

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.

American Psychological Association. (1974). *Standards for educational and psychological tests*. Washington, DC: Author.

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

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

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

Born, M. (2009, July). An ISO standard for assessment in work and organizational settings. In D. Bartram (Chair), *International guidelines and standards relating to tests and testing*. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

Buros, O. K. (1938). *The 1938 Mental Measurements Yearbook*. New Brunswick, NJ: Rutgers University Press.

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

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

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

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

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

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

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

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2009). *COTAN Beoordelingssysteem voor de Kwaliteit van Tests (geheel herziene versie)* [COTAN Rating system for test quality (completely revised edition)]. Amsterdam: NIP.

Evers, A., Zaal, J. N., & Evers, A. K. (2002). Ontwikkelingen in testgebruik over een periode van 33 jaar [Developments in test use over a period of 33 years]. *De Psycholoog*, *37*, 54–61.

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.

International Test Commission. (2003). *International Guidelines on Computer-Based and Internet Delivered Testing*. Bruxelles, Belgium: Author.

Keuning, J. (2004). *De ontwikkeling van een beoordelingssysteem voor het beoordelen van 'Computer Based Tests'* [The development of a rating system for the evaluation of 'Computer Based Tests']. POK Memorandum 2004–1. Arnhem, The Netherlands: Citogroep.

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.

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

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

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

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

NIP. (1969). *Documentatie van Tests en Testresearch in Nederland* [Documentation of Tests and Testresearch in The Netherlands]. Amsterdam: Nederlands Instituut van Psychologen.

Novick, M. R., & Lewis, C. (1987). Coefficient alpha and the reliability of composite measures. *Psychometrika*, *82*, 1–13.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer Verlag.

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

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

Resing, W. C. M., Evers, A., Koomen, H. M. Y., Pameijer, N. K., & Bleichrodt, N. (2008). Indicatiestelling speciaal onderwijs en leerlinggebonden financiering. Condities en instrumentarium [Indication for special indication and pupil-bound financing. Conditions and instruments]. Amsterdam: Boom test uitgevers.

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.

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

Sijtsma, K. (2009a). Correcting fallacies in validity, reliability, and classification. *International Journal of Testing*, *9*, 167–194.

Sijtsma, K. (2009b). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120.

Spies, R. A., Carlson, J. F., & Geisinger, K. F. (Eds.) (2010). *The eighteenth mental measurements yearbook*. Lincoln, NE: University of Nebraska Press.

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

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

Ter Laak, J. J. F., & De Goede, M. P. M. (2003). *Psychologische diagnostiek. Inhoudelijke en methodologische grondslagen* [Psychological diagnostics. Foundations regarding content and methodology]. Lisse, The Netherlands: Swets & Zeitlinger.

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

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* (pp. 59–76). Alphen aan de Rijn, The Netherlands: Samsom.

Wechsler, S. (2007). Test standards, development and use in Brazil. *Testing International*, *17*(June), 3–4.

Wools, S., Sanders, P., & Roelofs, E. (2007). *Beoordelingsinstrument: Kwaliteit van competentie assessment* [Evaluation instrument for the quality of competence assessment]. Arnhem, The Netherlands: Cito.

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

APPENDIX

1. Theoretical basis of the test

1.1. (Key question) Is the purpose of the test specified?

a. Is it described which construct(s) the test intends to measure?

b. Is (are) the group(s) for which the test is (are) intended specified?

c. Is the application of the test specified?

If the response to any of these three sub-items is negative, proceed to item 2.1.

1.2. Is the source of the construction idea been described, and/or is (are) the construct(s) that the test measures clearly defined?

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

2. Quality of the test materials, paper-and-pencil version

2.1. (Key question) Are the test items standardized?

2.2.a. (Key question) Is there an objective scoring system, and/or:

2.2.b. (Key question) If the test is being scored by raters or observers, is the system for assessment and observation clear and complete?

2.3. (Key question) Are the test items free from racist, ethnocentric, or sexist content or any other content offensive to specific groups of people?

If the response to any of the aforementioned items is negative, proceed to item 3.1.

2.4. Are the items, test booklet, answering scales, and answer form devised in a way that errors can be avoided when filling in?
2.5. Are the instructions for the subject of the test clear and complete?
2.6. Are the items correctly formulated?
2.7. How is the quality of the test materials?
2.8. Is the scoring system devised in such a way that errors can be avoided during scoring?

2. Quality of the test materials, computer version

2.9. (Key question) Are the test items standardized or are there explicit decision rules in the case of adaptive tests?
2.10. (Key question) Is scoring computerized or is there an objective scoring system?
2.11. (Key question) Are the test items free from racist, ethnocentric, or sexist content or any other content offensive to specific groups of people?
If the response to any of the aforementioned items is negative, proceed to item 3.1.
2.12. Is the software devised in such a way that errors caused by improper use can be avoided?
2.13. Are the instructions for the subject of the test clear and complete?
2.14. Are the items correctly formulated?
2.15. How is the quality of the user interface layout?
2.16. Is the test sufficiently secure?

3. Comprehensiveness of the test manual

3.1. (Key question) Is a test manual available?
If the response to this item is negative, proceed to item 4.1.
3.2. Are the instructions for the test administrator clear and complete?
3.3. Is information provided on user options and limitations of the test?
3.4. Is a summary of the research findings published in the manual?
3.5. Are case descriptions used to indicate how test scores could be interpreted?
3.6. Is there reference to types of information that could be significant for the interpretation of the test scores?
3.7. Is the degree of expertise required to administer and to interpret the test specified?

Additional questions for computer-based testing

3.8. Is information provided on installation of the computer software?
3.9. Is information provided on the operation of the software and the options it presents?
3.10. Is sufficient technical support provided?

4. Norms

4.1. (Key question) Are norms provided?
4.2. (Key question) Are the norms current?
If the response to any of the aforementioned items is negative, proceed to item 5.1.

Norm-referenced interpretation

4.3. (Key question) How is the quality of the supplied norm groups?
   a. Are the norm groups of sufficient size?
   b. Are the norm samples representative for the referred groups?

If the response to any of the aforementioned sub-items is negative, proceed to item 5.1.

4.4. Are the significance and the limitations of the norm scale used made clear to the user, and is the type of scale consistent with the objective of the test?
4.5. Is there information on means, standard deviations, and score distributions?
4.6 Is there information on possible differences between subgroups (for instance with respect to gender and ethnicity)?
4.7. Is there information on the accuracy of the measurements and the appropriate confidence intervals?
   a. Standard error of measurement, or
   b. Standard error of estimate, or
   c. Test information function/standard error.

Domain-referenced interpretation

4.8. Is there sufficient agreement between raters?
4.9. Are the procedures for determining the cut scores correct?
4.10. Have the raters been selected and trained appropriately?

Criterion-referenced interpretation

4.11. Do the research findings justify the use of cut scores?
4.12. Is the composition of the research group consistent with the intended purpose of the test?
4.13. Is the size of the research group sufficient?

5. Reliability

5.1. (Key question) Is information on the reliability of the test provided?

If the response to this item is negative, proceed to item 6.1.

5.2. Are the findings of the reliability research sufficient with respect to the intended type of decisions to be made with the aid of the test?
   a. Parallel-form reliability.
   b. Reliability based on inter-item relations.
   c. Test-retest reliability.
   d. Inter-rater reliability.
   e. Methods based on item-response theory.
   f. Methods based on generalizability theory or structural equation modelling.

5.3. 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?

6. Construct validity

6.1. (Key question) Is there information about the construct validity of the test?
If the response to this item is negative, proceed to item 7.1.
6.2. Do the findings of the validity research support the intended construct(s) being measured—or do the findings of the validity research make 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 potential item bias across subgroups.
d. Convergent and discriminant validity.
e. Differences in mean scores between relevant groups.
f. Other findings.

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

a. Are the procedures used in obtaining and computing data on construct validity correct?
b. Are the samples used in the research on construct validity consistent with groups for whom the test is intended?
c. What is the quality of the other measures used in the construct validity research?
d. Is the quality of the research, as rated in the items 6.3.a up to 6.3.c, good enough to corroborate the assessment of the construct validity as given in item 6.2?

7. Criterion validity

7.1. (Key question) Is there information about the test-criterion relationship?
If the response to this item is negative, items 7.2 and 7.3 can be skipped.
7.2. Are the findings of the validity research sufficient with respect to the type of decisions to be made with the test?
7.3. What is the quality of the criterion validity research?

a. Are the procedures used in collecting and calculating data on 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 the quality of the research, as rated in the items 7.3.a up to 7.3.c, good enough to corroborate the assessment of the criterion validity as given in item 7.2?