Chapter 4
Assessing Computer-Based Assessments

Bas Hemker, Cor Sluijter and Piet Sanders

Abstract Quality assurance systems for psychological and educational tests have been available for a long time. The original focus of most of these systems, be it standards, guidelines, or formal reviewing systems, was on psychological testing. As a result, these systems are not optimally suited to evaluate the quality of educational tests, especially exams. In this chapter, a formal generic reviewing system is presented that is specifically tailored to this purpose: the RCEC review system. After an introduction with an overview of some important standards, guidelines, and review systems, and their common backgrounds, the RCEC review system for the evaluation of educational tests and exams is described. The underlying principles and background of this review system are explained, as well as the reviewing procedure with its six criteria. Next, the system is applied to review the quality of a computer-based adaptive test: Cito's Math Entrance Test for Teachers Colleges. This is done to illustrate how the system operates in practice. The chapter ends with a discussion of the benefits and drawbacks of the RCEC review system.

4.1 Introduction

Quality assurance systems for psychological and educational tests have been available for a long time. These systems have their origins in the need to serve the public interest. They provide professional users with information to determine whether these instruments are suitable for the user's purpose. Quality assurance systems come in different forms. A common differentiation is between codes, guidelines, standards, and review systems (e.g., Roorda 2007). Codes are a cohesive set of behavioral rules with which test authors are expected to comply in order to make good and fair tests. As such, they are different from the other three as they do not reflect on the test

B. Hemker (✉) · C. Sluijter
Cito, Arnhem, The Netherlands
e-mail: bas.hemker@cito.nl

P. Sanders
RCEC, Vaassen, The Netherlands

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_4

itself. Guidelines are intended to show how a test should be developed. Standards are slightly different as they describe a level of quality that should be attained by a test on the aspects deemed relevant. Review systems critically evaluate a psychological or educational test in order to make it possible to decide whether or not it has sufficient fit to purpose.

The first document containing a set of systematic evaluations of tests was the 1938 Mental Measurements Yearbook (Buros 1938). This volume contained a set of critical reviews of various psychological tests, questionnaires, and rating scales then in use. It was intended to assist professionals to select and use the most appropriate psychological test for their specific problem. Spanning a period of almost eight decades, its twentieth edition was published in 2017 (Carlson et al. 2017). Nowadays, the system used to review all instruments in the Mental Measurements Yearbook is accompanied by a profusion of other quality assurance systems.

Well-known guidelines on the development and use of psychological tests were developed and are maintained by the International Test Commission (ITC). This is an association of national associations of psychologists, test committees, and other organizations and individuals promoting the proper development and use of tests. The ITC came into existence in the 1970s (see, Oakland et al. 2001). The ITC now oversees six different guidelines, including guidelines for test use (ITC 2013), test security (ITC 2014), and computer-based and internet testing (ITC 2005).

The best known standards to date are the Standards for Educational and Psychological Testing. These standards have been jointly published in different editions since 1966 by three institutes: the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The most recent publication dates from 2014 (APA et al. 2014). The different editions were preceded by separate documents called "Technical recommendations for psychological tests and diagnostic techniques" (APA 1954) and "Technical recommendations for achievement tests" (AERA and NCME 1955). These publications primarily addressed the developmental process of tests as well as the type of information publishers should make available to test users in manuals (Camara and Lane 2006).

The European Federation of Psychologists' Associations (EFPA) developed and maintains a review system for the description and evaluation of psychological tests. Its development started in 1989, but the first formal version was published much later (Bartram 2002). The most recent version was issued in 2013 (EFPA 2013). The current system was greatly influenced by criteria used earlier by the British Psychological Society and the Dutch Committee on Testing (COTAN) of the Dutch Association of Psychologists. This latter body, which was founded in 1959, itself has a long tradition in evaluating tests. COTAN started publishing test reviews in 1969 (Nederlands Instituut van Psychologen 1969). The most recent revision of the system was published almost a decade ago (Evers et al. 2010a, 2010b). In 2019 COTAN initiated work on a new revision. For a detailed overview of the history of the introduction of review models for psychological tests in Europe, the interested reader is referred to Evers (2012).

An analysis of the content of these and other current quality assurance systems, for instance, those of the Educational Testing Service (2014), the Association for Educational Assessment Europe (2012), and Cambridge Assessment (2017), demonstrates that all systems show substantial overlap. This is not surprising, because they all reflect in one way or another the theoretical and practical developments in psychological and educational measurement from the first half of the last century up until now, as captured in, for instance, the consecutive editions of Educational Measurement (Brennan 2006; Lindquist 1951; Linn 1989; Thorndike 1971). Unfortunately, all these systems also have a basic flaw that comes forth from their origins. Because they all had psychological tests as their original focus, the evaluation of educational tests, and especially exams as a specific subset, raises several issues

An exam can be defined as a formal investigation by a licensed body into the knowledge base, abilities, attitudes, skills and/or competencies of candidates. In order to receive a diploma or a certificate, candidates have to demonstrate a certain level of mastery on a set of assignments that are representative of the total domain of assignments for (certain parts of) the corresponding curriculum. This definition points out a generic difference between psychological and educational tests: the construct that is intended to be measured.

Educational tests are more often than not direct measures of human behavior, such as mathematical ability, reading comprehension, or spelling ability, while the constructs that are the subject of psychological tests have a more theoretical character like intelligence or neuroticism. This has as a direct consequence that criteria having to do with this aspect of test validity have a different orientation and weight for educational and psychological tests.

Another difference is that for most psychological constructs, some sort of stability over time is assumed, whereas the constructs measured in education are subject to change. While students are learning, they are expected to change over time in their ability—hopefully increasing after more education. On the other hand, the ability may also decrease, for example, due to a lack of practice, or simply forgetting things that once were learned. What is being measured is often a snapshot in time. The temporality is also reflected by the single use of many exams: they are used once, at one moment in time, never to be used again for the same purpose.

A lack of stability is also reflected in the constructs themselves as they can change over time as well: what was considered relevant in mathematics can change over time and over levels of ability. On the other hand, even if educational tests and exams change over time, regulators want to compare results over time. This means that in comparison with psychological tests, equating procedures to maintain stable norms are especially important. All this has consequences for the way tests and exams are reviewed.

4.2 The RCEC Review System for the Evaluation of Computer-Based Tests

The Research Center for Examinations and Certification (RCEC) has developed an analytical review system that is specifically tailored to evaluating the quality of educational tests, and particularly exams (Sanders et al. 2016). It was in large part inspired by the aforementioned COTAN review system. An overview of the principles and background of the RCEC review system is presented below, including the description of the six criteria the system uses.

The RCEC review system has three main characteristics in common with other review systems such as the EFPA system and the COTAN system. First, it focusses on the intrinsic quality of the instrument itself and not on the process of test development. It evaluates the quality of items and tests, but not the way they are produced. Of course, the scientific underpinning of the test reveals much about the process of test development, but this is only reviewed in the light of the impact of the process on the quality of the items and the test as a whole. Secondly, the review system works with a set of criteria with which an educational test should comply in order to be considered to be of sufficient quality. The third characteristic is that the review is completely analytical, or can even be considered to be actuarial. For each criterion, the reviewer answers a series of questions by giving a rating on a three-point scale: insufficient—sufficient—good. The review system contains clarifications of the questions. Instructions are provided to ensure that the scores ensuing from these questions are as objective as possible. For each of the six criteria the system applies, the ratings are combined through specific rules that yield a final assessment of the quality of each criterion, again on this three-point scale.

The nature of the selected criteria and their specific description is where the RCEC review system differs from others. Other systems, like the EFPA system and originally the COTAN system as well, focus more on psychological tests. As already mentioned, other criteria apply, or have a different weight when it comes to educational tests and especially exams. The RCEC review system consequently differs from other systems in the wording of the criteria, the underlying questions, their clarifications, the instructions, and the scoring rules. This is done in order to have the best fit for purpose, i.e., the evaluation of educational assessment instruments and exams in particular.

The reviewing procedure shares some features with other reviewing systems. Like the Buros, EFPA, and COTAN systems, two reviewers evaluate an educational test or exam independently. The reviewers are non-anonymous. Only professionals who are certified by RCEC after having completed a training in using the review system are allowed to use it for certification purposes. Note that all three authors of this chapter have this certificate. All cases are reviewed by the overseeing Board. They formulate the final verdict based on the advice of the reviewers.

The criteria of the RCEC system are:

- Purpose and use;
- Quality of test and examination material;

- Representativeness;
- Reliability;
- Standard setting, norms, and equating;
- Administration and security.

There is overlap with the criteria of other quality assurance systems. 'Purpose and use', 'Quality of test and examination material', 'Reliability', and 'Administration and security' can also be found in other systems. The most notable difference between the RCEC review system and other systems rests in the criterion of 'Representativeness' which corresponds with what other systems refer to as (construct and criterion-related) validity, but uses a different approach, especially for reviewing exams. Since these are direct measures of behavior rather than measures of constructs, the focus of this criterion is on exam content. Another difference is that within the criterion of 'Standard setting, norms, and equating', more attention is given to the way comparability over parallel instruments is ensured. It details how equivalent standards are being set and maintained for different test or exam versions.

Below, the criterion 'Purpose and use' is discussed in detail. This criterion is emphasized, because it is often taken for granted. Its importance cannot be overstated, as in order to produce a quality educational test or exam, it is vital that its purpose is well-defined. For the other five criteria, a shorter overview is given. Similar to the first criterion, these criteria are also found in other review systems. In this overview, special attention is given to the criteria as applied to computerized tests. This is done because the application of the review system is demonstrated by the evaluation of the quality of a computerized adaptive test (CAT).

A detailed description of the whole RCEC review system can be found at www.rcec.nl. Currently, the review system is only available in Dutch. An English version is planned.

4.2.1 Purpose and Use of the Educational Test or Exam

The golden rule is that a good educational test should have one purpose and use only. The exception to this is a situation where different purposes are aligned. For instance, a formative test can help in making simultaneously decisions on an individual, group, or school level, simultaneously. Discordant purposes and uses (e.g., teacher evaluation versus formative student evaluation) should not be pursued with one and the same educational test. This would lead to unintended negative side effects. In most cases, the purpose of educational tests and exams is to assess whether candidates have enough knowledge, skills, or the right attitudes. The use of an educational test concerns the decisions that are made based on the score obtained.

There are three questions used to score a test on this criterion:

- Question 1.1: Is the target population specified?
- Question 1.2: Is the measurement purpose specified?
- Question 1.3: Is the measurement use specified?

Question 1.1. has to do with the level of detail in the description of the test or exam target groups(s). Age, profession, required prior knowledge, and the level of education can also be used to define the target group. Without this information, the evaluation of the language used in the instructions, the items, the norm, or cut scores of the test becomes troublesome. Question 1.1 relates to who is tested and when. A test or exam gets a rating 'Insufficient' (and a score of 1) for this question when the target group is not described at all, or not thoroughly enough. This rating is also obtained when the educational program of studies for the target group is not described. A test gets a rating 'Sufficient' (a score of 2) only when the educational program the test is being used for is stated. It receives a rating 'Good' (a score of 3) for this question if not only the educational program but also other relevant information about the candidates is reported. This detailed information includes instructions on the application of the test to special groups, such as students having problems with sight or hearing.

An educational test should assess what candidates master after having received training or instruction. This is what question 1.2 refers to. What candidates are supposed to master can be specified as mastery of a construct (e.g., reading skill); of one of the subjects in a high school curriculum (e.g., mathematics); of a (component of a) professional job; or of a competency (e.g., analytical skills in a certain domain). A test that measures a construct or a competency needs to present a detailed description with examples of the theory on which the construct or competency is based. This implies that tautological descriptions like 'this test measures the construct reading skills' do not suffice. The construct or competency has to be described in detail and/or references to underlying documents have to be presented. The relevance of the content of the test or exam for its intended purpose should be clarified. A blueprint of the test can be a useful tool in this regard. A rating 'Insufficient' is given when the measurement purpose of the test is not reported. A rating 'Sufficient' is given when the purpose is reported. A rating 'Good' is given when in addition to this, a (detailed) description of constructs, competencies, or exam components is supplied as described above.

Educational tests or exams can be used in many ways. Each use refers to the type of decision that is being made based on the results of the test(s) and the impact on the candidate. Common uses are selection or admittance (acceptance or refusal), classification (different study programs resulting in different certificates or degrees), placement (different curricula that will result in the same certificate or degree), certification (candidates do or do not master a certain professional set of skills), or monitoring (assessment of the progress of the candidates). Question 1.3. is dichotomously scored: either the use of the test is reported in enough detail ('Good'), or it is not ('Insufficient').

The overall evaluation of the description of the purpose and use of the test is based on the combination of scores on the three questions. The definite qualification for this criterion is 'Good' if a test receives a score of 3 on all three questions, or if two questions have a score 3 while the third one a score of 2. If Question 1.3 is scored 3 and the other two are scored 2, the qualification 'Sufficient' is given. Finally, the

qualification is 'Insufficient' if one of the three questions was awarded a score of 1. This means that all three items are knock-out questions.

4.2.2 Quality of Test Material

All test material (manual, instructions, design, and format of items, layout of the test, etc.) must have the required quality. The items and the scoring procedures (keys, marking scheme) should be well defined and described in enough detail. The same holds for the conditions under which the test is to be administered.

The following key questions are considered:

- Question 2.1: Are the questions standardized?
- Question 2.2: Is an objective scoring system being used?
- Question 2.3: Is incorrect use of the test prevented?
- Question 2.4: Are the instructions for the candidate complete and clear?
- Question 2.5: Are the items correctly formulated?
- Question 2.6: What is the quality of the design of the test?

The first two questions are knock-out questions. If on either one of the two, a score of 1 is given, the criterion is rated 'Insufficient' for the test.

The RCEC review system makes a distinction between paper-and-pencil tests and computer-based tests. Some remarks on the application of the system for a CAT can be made. First, the next item in a CAT should be presented swiftly after the response to the previous item(s). In evaluating a CAT, Question 2.2 implies that there should be an automated scoring procedure. Secondly, Question 2.3 implies that software for a CAT should be developed such that incorrect use can be prevented. As the routing of the students through the test depends on previously given answers, going back to an earlier item and changing the response poses a problem in a CAT. Finally, Question 2.6 refers to the user interface of the computerized test.

4.2.3 Representativeness

Representativeness relates to the content and the difficulty of the test or exam. This criterion basically refers to the content validity of the test: do the items or does the test as a whole reflect the construct that is defined in Question 1.2. The key question here is whether the test (i.e., the items it contains) is actually measuring the knowledge, ability, or skills it is intended to measure. This can be verified by the relationship between the items and the construct, namely, the content. This criterion is evaluated through two knock-out questions:

- Question 3.1: Is the blueprint, test program, competency profile, or the operationalization of the construct an adequate representation of the measurement purpose?

- Question 3.2: Is the difficulty of the items adjusted to the target group?

Note that this criterion has a structurally different approach compared to corresponding criteria from review systems with their focus on psychological tests. Question 3.1 specifically refers to the content of a test or exam: it should be based on what a candidate has been taught, i.e., learning objectives. As these learning objectives often are not specific enough on which to base the construction of a test, classification schemes, or taxonomies of human behavior are used to transform the intended learning objectives to objectives that can be tested. Since educational tests, and especially exams are generally direct measures of behavior rather than measures of constructs, priority is given here to the content of the test or exam. In a CAT this also means that extra constraints have to hold to assure that candidates get the appropriate number of items for each relevant subdomain.

Question 3.2 asks whether the difficulty of the items, and thus the difficulty of the test or exam, has to be adjusted to the target group. In practice, this means that a test should not be too difficult or too easy. Particularly in a CAT, where the difficulty of the question presented is targeted to the individual taking the test, this should be no problem. The only issue here is that there should be enough questions for each level of difficulty.

4.2.4 Reliability

The previous two earlier criteria focus mainly on the quality of the test items. The evaluation of reliability involves the test as a whole. It refers to the confidence one can have in the scores obtained by the candidates. Reliability of a test can be quantified with a (local) reliability coefficient, the standard error of measurement, or the proportion of misclassifications. The first of the three questions is a knock-out question:

- Question 4.1: Is information on the reliability of the test provided?
- Question 4.2: Is the reliability of the test correctly calculated?
- Question 4.3: Is the reliability sufficient, considering the decisions that have to be based on the test.

In the case of a CAT, traditional measures for reliability do not apply. A CAT focusses on minimizing the standard error of measurement by following an algorithm that sequentially selects items that maximize the statistical information on the ability of the candidate, taking into consideration a set of constraints. The information function drives the selection of items, and the evaluation of the standard error of measurement is one of the important criteria to stop or to continue testing. Thus, without a positive answer on question 4.1, a CAT is not possible. Question 4.3 can be interpreted in a CAT by checking whether the stopping rule is appropriate given the purpose and use of the test, and whether there are sufficient items to achieve this goal.

4.2.5 Standard Setting and Standard Maintenance

This criterion reviews the procedures used to determine the norms of a test, as well as how the norms of comparable or parallel tests of exams are maintained. Norms can be either relative or absolute. If the norms were previously determined but need to be transferred to other tests or exams, equivalence and equating procedures need to be of sufficient quality. There are separate questions for tests or exams with absolute or relative norms.
Questions for tests with absolute norms:

- Question 5.1: Is a (performance) standard provided?
- Question 5.2a: Is the standard-setting procedure correctly performed?
- Question 5.2b: Are the standard-setting specialists properly selected and trained?
- Question 5.2c: Is there sufficient agreement among the specialists?

Questions for tests with relative norms:

- Question 5.3: Is the quality of the norms sufficient?
- Question 5.3a: Is the norm group large enough?
- Question 5.3b: Is the norm group representative?
- Question 5.4: Are the meaning and the limitations of the norm scale made clear to the user and is the norm scale in accordance with the purpose of the test?
- Question 5.5a: Is the mean and standard deviation of the score distribution provided?
- Question 5.5b: Is information on the accuracy of the test and the corresponding intervals (standard error of measurement, standard error of estimation, test information) provided?

Questions for maintaining standards or norms:

- Question 5.6: Are standards or norms maintained?
- Question 5.6a.: Is the method for maintaining standards or norms correctly applied?

A CAT can have absolute or relative norms, depending on the purpose and use of the test. However, for a CAT, the evaluation of the way the standards or norms are maintained most definitely needs to be answered, as each individual candidate gets his or her unique test. It is mandatory that the results from these different tests are comparable in order to make fair decisions. In CAT, this equating is done through item response theory (IRT). Question 5.6a relates to whether IRT procedures have been applied correctly in the CAT that is being reviewed.

4.2.6 Test Administration and Security

Information on how to administer the test or exam and how to assure a secure administration should be available for the proctor. The key concern is whether the design

of the test is described in such a way that, in practice, testing can take place under standardized conditions, and whether enough measures are taken to prevent fraud. The questions for this criterion are:

- Question 6.1: Is sufficient information on the administration of the test available for the proctor?
- Question 6.1a: Is the information for the proctor complete and clear?
- Question 6.1b: Is information on the degree of expertise required to administer the test available?
- Question 6.2: Is the test sufficiently secured?
- Question 6.3: Is information on the installation of the computer software provided?
- Question 6.4: Is information on the operation and the possibilities of the software provided?
- Question 6.5: Are there sufficient possibilities for technical support?

Question 6.1 refers to a proper description of what is allowed during the test. Question 6.2 refers to the security of the content (e.g., for most practical purposes, it should not be possible for a candidate to obtain the items before the test administration), but also refers to preventing fraud during the test. Finally, security measures should be in place to prevent candidates altering their scores after the test is administered.

This means that it should be clear to a test supervisor what candidates are allowed to do during the administration of a CAT. In order to get a 'Good' on this criterion, it must be made clear, for example, whether the use of calculators, dictionaries, or other aids is allowed in the exam, what kind of help is allowed, and how to handle questions from the examinees. The security of CAT is also very much dependent on the size and quality of the item bank. A CAT needs measures to evaluate the exposure rate of items in its bank. Preferably, measures for item parameter drift should also be provided.

4.3 Reviewing a Computer Based Test

The usefulness of a review system is best demonstrated by its application. Therefore, Cito's Math Entrance Test for Teachers College (WISCAT-pabo) is evaluated below with the RCEC review system. This was done by two independent certified reviewers, who did not differ as far as the ratings on all criteria are concerned. The WISCAT-pabo is a compulsory high stakes CAT in the Netherlands, developed by Cito. It is a test of arithmetic for students in their first year of primary school teacher education. The test has been in use for over a decade with regular updates of its item bank. Candidates get three attempts to score above the cut score. If they fail the test, they cannot continue their teacher training. The psychometric advantages of computer-based adaptive testing in this instance are obvious: efficiency, high measurement precision, and prevention of the test content becoming exposed. The instrument is reviewed separately for each criterion. The review is for the most part based on a number of sources: information from the WISCAT-pabo manual (Straetmans and

Eggen 2007) that contains a detailed technical report, information on the Cito website (https://www.cito.nl/onderwijs/hoger-onderwijs/ho-toetsen-pabo/wiscat-pabo), and the reviewers taking the test several times.

4.3.1 Purpose and Use of the Test

The target population is well defined, consisting of incoming and first-year teachers college students. However, the manual does not provide information on whether or not the instrument is also suited for students with special needs. The test can be taken at various moments, but these are limited in number to assure security of the item bank.

The purpose of the test is defined as measuring the level of calculation skills of incoming and first-year students. A very detailed description of the construct of calculation skills is provided. Calculation skills are described for four domains for four different math levels: (1) Calculations and measures; (2) Geometry; (3) Information processing, statistics, and probability; and (4) Algebra, connections, graphs, and functions. The test tackles basic skills (counting, addition, subtraction, multiplication, division, powers, estimates, rounding), fractions, percentages, ratios, decimal numbers, among others. Within these domains, 21 subdomains are given and 178 descriptions of subskills.

Finally, the intended use of the WISCAT-pabo is extensively discussed in the manual. The instrument is intended to determine whether incoming students have sufficient arithmetical knowledge and skills to successfully develop the subject-specific and subject-didactic knowledge and skills to a level that is required to learn how to teach arithmetic to pupils in primary education. In addition, the instrument can serve a formative purpose when a candidate scores below the cut score. It then provides global indications on the level of mastery of several subdomains, making it possible for students to specifically focus on the subdomains they master the least.

The review yields full points for Questions 1.2 and 1.3. The score on Question 1.1 is 2 (‘Sufficient’), because very limited information was provided on the use of the test for special groups. The scoring rules yield ‘Good’ as the overall score for this criterion.

4.3.2 Quality of Test Material

The items are standardized. Each student gets a set of 50 items. The items are selected from a database of over 900 items. There are multiple-choice questions and short open-ended questions that are automatically scored. Items answered correctly yield a score of 1; items answered incorrectly get a score of 0. Both an overall score on the ability scale and indicative profile scores are generated. Because of the nature of the CAT, it is not possible for candidates to review earlier items. This is often

seen as a disadvantage and is one of the reasons other methods of testing, such as multistage-testing are becoming more popular to use in high-stakes testing (van Boxel and Eggen 2017).

The instructions for the candidates are complete and clear. The Cito website provides a well-written six-page instruction. Not all 900 items in the item bank were evaluated for this review, but all items were developed by experienced item writers well acquainted with the specific subject matter. The quality of the content of the items as well as their psychometric quality is guaranteed. Items were developed through an established procedure in which experienced test developers thoroughly checked the items and all items were piloted with potential test takers. The psychometric evaluation took place through pretest procedures and continuous monitoring of new incoming data, Finally, the quality of the interface of the WISCAT-pabo can also be rated as 'Good'. Based on all the ratings on the questions and the application of the scoring rules, the instrument receives the rating 'Good' for this criterion.

4.3.3 Representativeness

A well-designed educational test or exam reflects the objectives of the (part of the) curriculum it is intended to measure. To achieve this, a test matrix is drawn up early in the design phase. This can be considered the blueprint of the test in which the test material is depicted by two dimensions, respectively the operations that students must be able to carry out and the subject matter. The test items must be evenly distributed over both dimensions.

In a CAT, the composition of the test primarily takes into account the reconciliation of the difficulty of the items with the provisional estimate of the student's skill. Without special measures, the computer will not pay attention to the distribution of the items on the subject matter and the operations. A straightforward way to guarantee this is the design of a constrained CAT in which the item bank is compartmentalized. From each part of this item bank, a specified minimum number of items is selected. A sufficient number of items needs to be available in each part of the bank, thus the developers of the CAT need to provide a balanced item bank. Otherwise, the algorithm does not produce tests that are representative for the relevant subdomains.

As the WISCAT-Pabo must generate tests that, in addition to an overall ability estimate, also provide an indication of the level of mastery of four subdomains, the CAT is designed in such a way that sufficient items from each subdomain are selected. In the WISCAT-Pabo, 50 items are presented to the students, with a very strict distribution over subdomains. Thus, all subdomains are covered. As this is a CAT and items are selected and presented based on the estimated ability of the candidate, the item difficulty is by definition at the right level for the candidates. Note that this optimal selection depends on the availability of sufficient items on the whole range of relevant abilities. The main purpose of the test is to check whether the candidates pass a specifically set cut-off score. It turns out that given the ability of the candidates this is somewhat in the middle of the ability range. Therefore it may

not be too much of an issue whether there are enough items available for test takers of very high or very low ability. Also a procedure for exposure control is applied, and in case of over exposure items are to be replaced by new equivalent items. With over 900 items in the initial bank (18 times the test size), the distribution is also covered well. With maximum scores on all questions the review for this criterion results in a 'Good'.

4.3.4 Reliability (Measurement Precision)

Because this is a CAT that uses the item and test information at the heart of its procedure, measurement precision is definitely provided. Experience shows that with traditional paper-and-pencil calculation skills tests, a total number of 50 items yields high reliability. Additionally, early research has shown that CATs measure just as accurately with about half the number of items as traditional paper-and-pencil tests (Vispoel et al. 1994). The actual reduction depends on the composition of the item bank and the item selection criterion.

Thus, in the case of the WISCAT-pabo, the reliability is good. Studies performed by the authors also confirm this: the estimated mean reliability for each test is 0.91. This result was found both in simulation studies as well as in operational results. The manual provides all necessary formulas, based on relevant literature, such as Thissen (2000). It can be concluded that the way reliability was calculated is correct. In addition, the authors also provide the percentages of correct and incorrect pass-fail decisions, based on a simulation study. They show that the percentage of correct decisions is 91.54%, with about an equal percentage of candidates incorrectly passing or failing. These are good percentages, considering the passing rate of about 45%. With the additional opportunity to do a resit of the test, the number of students that fail the test while having sufficient ability is extremely small (about a fifth of a percent, after one resit). It is almost nonexistent after two resits. The review of this criterion therefore also results in a 'Good'.

4.3.5 Standard Setting and Standard Maintenance

The norms for the WISCAT-pabo are set by a combination of procedures, mainly relative. The cut score is set on the ability scale at the position of the 80th percentile of the grade 8 student population, the last grade in primary education. The underlying rationale is that based on experts' opinion, the ability of a student starting teacher training should at least be at the level of a good grade 8 student. A good grade 8 student for calculation skills was next defined at the minimum level for the top 20% of grade 8 students. This corresponding cut score on the ability scale of the WISCAT-Pabo was determined by an equating study relating the math results of 150,722 grade 8 students in 2002 on the End of Primary Education test to the results

on the WISCAT-pabo. This size is most definitely large enough and, with over 87% of the total population included, it is also representative. The equating study used the OPLM model (Verhelst and Eggen 1989; Verhelst and Glas 1995), a variant of the two-parameter logistic test model. The design of the study was described in detail, as well as the meaning and the limitations of the norm scale. The results were related to the results of a sample of 4805 aspiring teachers. A wide variety of distribution characteristics was given, including the distribution of the ability of these aspiring teachers. As IRT takes such a crucial role in the procedure to set the norm, and as IRT is also crucial in the application of a CAT, it is obvious that IRT was used to maintain the standards. All procedures that were described were correctly applied. Rating all the questions in total, the review for this criterion results in a 'Good'.

4.3.6 Test Administration and Security

Most relevant information on the administration of the test is available on the Cito website. This includes the information on installing the computer software, the way the software operates, and possibilities for technical support. The safety measures include an aspect of the CAT algorithm which prevents the over- and under-utilization of items in the bank. Simply put, before a new test is started, part of the data bank is shielded from the test algorithm (counteracting overuse). The selection of items is based on a mixture of strictly adaptive and strictly random selection, while the relationship between the two modes shifts in the direction of adaptive selection with each successive item. This procedure can lead to a candidate being given an item, sometimes at the beginning of the test, which is obviously far too difficult or too easy, based on the currently estimated skills of that candidate. More detailed references, e.g., Sympson and Hetter (1985), and Revuelta and Ponsoda (1998), are given in the manual. Reviewing the responses to the questions, and the scoring rules of the review system, this criterion also yielded a rating 'Good'.

4.3.7 Review Conclusion

As the review for all criteria was positive, the conclusion is that the WISCAT-Pabo is fit for its purpose. Thus it can be used by Teachers Colleges in the Netherlands to decide whether or not starting students have sufficient calculation skills to continue their training. In addition, the WISCAT-Pabo can be used in a formative way by students scoring below the cut score to find out if there are specific subdomains in arithmetic that they should focus on.

4.4 Discussion

In the previous sections, the RCEC review system was described. Also an example of a specific review was presented. We would like to stress here that this was not a formal review and that it was only performed to make clear that reviewing the quality of educational tests and exams requires a structurally different approach than reviewing the quality of psychological tests. The field of educational measurement is still developing and improving. This means that the RCEC review system will have to be updated on a regular basis.

We hope that the RCEC review system will enjoy increasing popularity within the educational measurement community. One of the reasons is that the RCEC review system is designed to deal with one of the principal differences between exams and psychological tests. The content of the latter can remain the same over an expanded period of time, whereas the content of most exams can be deployed only once. The reason for this, of course, is that, in high-stake situations, exam content becomes known or is even made public after the first administration. This exposure makes it impossible for candidates to do a resit of the exam since future candidates can become directly acquainted with the exam content. Again, this has consequences for the application and weight of criteria like ageing of research findings. For exams, the issue of equating is much more relevant than updating the norm because it is outdated. The lifespan of a specific educational test is simply too short for that to happen.

Another reason we see for its increasing popularity is that the RCEC review system has a generic nature. It can be further adapted for reviewing different types of educational tests, including multiple-choice exams, open-ended exams, oral exams, performance exams, or, as we have shown here, computer-based educational tests. Furthermore, we would like to make a plea to not only use this system for reviewing educational tests and exams, but also as a concrete guideline for producing educational tests and exams. This could help inexperienced test and exam developers to increase their level of expertise in an efficient way, thus increasing the quality of the instruments they produce.

References

American Educational Research Association, & National Council on Measurement Used in Education. (1955). *Technical recommendations for achievement tests*. Washington, DC: National Educational Association.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: American Psychological Association.

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

Association of Educational Assessment—Europe. (2012). *European framework of standards for educational assessment 1.0*. Roma: Edizione Nova Cultura. Retrieved from www.

aea-europe.net, https://www.aea-europe.net/wp-content/uploads/2017/07/SW_Framework_of_European_Standards.pdf

Bartram, D. (2002). *Review model for the description and evaluation of psychological tests*. Brussels, Belgium: European Federation of Psychologists' Associations.

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: American Council on Education and Praeger Publishers.

Buros, O. K. (Ed.). (1938). *The 1938 mental measurements yearbook*. Oxford, England: Rutgers University Press.

Camara, W. L., & Lane, S. (2006). A historical perspective and current views on the standards for educational and psychological testing. *Educational Measurement: Issues and Practice, 25*(3), 35–41. https://doi.org/10.1111/j.1745-3992.2006.00066.x.

Cambridge Assessment. (2017). *The Cambridge Approach to Assessment. Principles for designing, administering and evaluating assessment* (Revised Edition). Retrieved from http://www.cambridgeassessment.org.uk/Images/cambridge-approach-to-assessment.pdf.

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.). (2017). *The twentieth mental measurements yearbook*. Lincoln, NE: Buros Center for Testing.

Cito. (2018). *Toetsen voor de pabo - Rekentoets wiscat* [Tests for Teachers College—The Wiscat Math Entrance Test]. Retrieved from https://www.cito.nl/onderwijs/hoger-onderwijs/ho-toetsen-pabo/wiscat-pabo.

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Princeton: Educational Testing Service.

European Federation of Psychologists' Associations. (2013). *EFPA review model for the description and evaluation of psychological and educational tests. Test review form and notes for reviewers. Version 4.2.6*. Retrieved from http://www.efpa.eu/professional-development/assessment.

Evers, A. (2012). The internationalization of test reviewing: Trends, differences and results. *International Journal of Testing, 12*(2), 136–156. https://doi.org/10.1080/15305058.2012.658932.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, S. (2010a). *COTAN Beoordelingssysteem voor de kwaliteit van tests*. [COTAN Review system for the quality of tests]. Amsterdam: NIP. Retrieved from www.psynip.nl, https://www.psynip.nl/wp-content/uploads/2016/07/COTAN-Beoordelingssysteem-2010.pdf.

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010b). The Dutch review process for evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing, 10*(4), 295–317. https://doi.org/10.1080/15305058.2010.518325.

International Test Commission. (2005). *International guidelines on computer-based and internet delivered testing*. Retrieved from www.intestcom.org, https://www.intestcom.org/files/guideline_computer_based_testing.pdf.

International Test Commission. (2013). *ITC guidelines on test use. Version 1.2*. Retrieved from www.intestcom.org, https://www.intestcom.org/files/guideline_test_use.pdf.

International Test Commission. (2014). *International guidelines on the security of tests, examinations, and other assessments*. Retrieved from www.intestcom.org, https://www.intestcom.org/files/guideline_test_security.pdf.

Lindquist, E. F. (Ed.). (1951). *Educational measurement*. Washington, D.C.: The American Council on Education.

Linn, R. L. (Ed.). (1989). *Educational measurement* (3rd ed.). New York, NY, England: Macmillan Publishing Co, Inc.: The American Council on Education.

Nederlands Instituut van Psychologen. (1969). *Documentatie van tests en testresearch in Nederland* [*Documentation of tests and test research in the Netherlands*]. Amsterdam: Nederlands Instituut van Psychologen.

Oakland, T., Poortinga, Y. H., Schlegel, J., & Hambleton, R. K. (2001). International Test Commission: Its history, current status, and future directions. *International Journal of Testing, 1*(1), 3–32. https://doi.org/10.1207/S15327574IJT0101_2.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(4), 311–327. https://doi.org/10.1111/j.1745-3984.1998.tb00541.x.

Roorda, M. (2007). Quality systems for testing. In R. Ramaswamy & C. Wild (Eds.), *Improving testing: Process tools and techniques to assure quality* (pp. 145–176). London: Routledge.

Sanders, P. F., van Dijk, P., Eggen, T., den Otter, D., & Veldkamp, B. (2016). *RCEC Beoordelingssysteem voor de kwaliteit van studietoetsen en examens*. [Review system for the quality of educational tests and exams]. Enschede: RCEC.

Straetmans, G., & Eggen, T. (2007). *WISCAT-pabo Toetshandleiding*. [WISCAT-pabo Test manual]. Arnhem: Cito.

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item-exposure rates in computerized adaptive testing*. Paper presented at the annual conference of the Military Testing Association, San Diego.

Thissen, D. (2000). Reliability and measurement precision. In: Wainer, H. (Ed.), *Computerized adaptive testing. A primer* (pp. 159–184). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thorndike, R. L. (Ed.). (1971). *Educational measurement* (2nd ed.). Washington, D.C.: The American Council on Education.

van Boxel, M., & Eggen, T. (2017). *The Implementation of Nationwide High Stakes Computerized (adaptive) Testing in the Netherlands.* Paper presented at the 2017 conference of the International Association for Computerised Adaptive Testing, Niigata, Japan. Retrieved from http://iacat.org/implementation-nationwide-high-stakes-computerized-adaptive-testing-netherlands-0.

Verhelst, N. D., & Eggen, T. J. H. M. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek.* [Psychometrical and statistical features of national assessment of educational progress] (PPON-rapport, nr. 4). Arnhem: Cito Instituut voor Toetsontwikkeling.

Verhelst, N. D., & Glas, C. A. W. (1995). The one-parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.). *Rasch models. Foundations, recent developments and applications* (pp 215–238). New York: Springer.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Applied Measurement in Education, 7*(1), 53–79. https://doi.org/10.1207/s15324818ame0701_5.