INTERNATIONAL TEST COMMISSION

ITC Guidelines for Quality Control in Scoring, Test Analysis, and Reporting of Test Scores

FINAL VERSION

May 2011

Please cite as:

INTERNATIONAL TEST COMMISSION (2011). ITC GUIDELINES FOR QUALITY CONTROL IN SCORING, TEST ANALYSIS, AND REPORTING OF TEST SCORES. [HTTP://WWW.INTESTCOM.ORG]

ACKNOWLEDGEMENTS

The Guidelines were prepared for the ITC Council by Avi Allalouf. The author is grateful to Marise Born for her valuable assistance in the execution of this project, as well as to the following individuals who reviewed previous versions and whose input was important to the development of the present Guidelines:

Alvaro Arce-Ferrer	Pearson Vue, USA
James Austin	Ohio State University, USA
Jo-Anne Baird	University Bristol, UK
Giulia Balboni	University of Valle d'Aosta, Italy
Helen Baron	Independent Consultant, UK
Dave Bartram	SHL, UK
James Butcher	University of Minnesota, USA
Janet Carlson	Buros Center for Testing, USA
Iain Coyne	University of Nottingham, UK
Kurt Geisinger	University of Nebraska-Lincoln, USA
Ron Hambleton	University of Massachusetts, USA
John Hattie	University of Auckland, New Zealand
Fred Leong	Michigan State University, USA
Jason Lamprinou	European University, Cyprus
Tom Oakland	University of Florida, USA
Fred Oswald	Rice University, USA
Christopher Rush	Wireless Generation, Inc., USA

The author is also indebted to his colleagues at NITE (National Institute for Testing and Evaluation), especially in the Scoring and Equating Department where many quality control procedures were developed and are routinely employed.

Contents

1. INTRODUCTION

1.1 Aim and objectives

Standardization and accuracy are essential in all stages of testing, beginning with test development and test administration, right through to scoring, test analysis, score interpretation and score reporting. Anyone involved in scoring, test analysis, and score reporting has a responsibility to maintain professional standards that can be justified to relevant stakeholders, including hiring organizations, psychological associations, colleges and universities, governing agencies and legal entities. The professional practitioner should be aware of, and anticipate, errors that can occur at any stage, and must act in accordance with current standards to prevent and address such errors. Inaccurate scores resulting from a wrong answer key, incorrect conversion of raw scores to standard scores, a mistake in the computation of a test record, accidental reporting of scores to the wrong client, or misinterpretation of the reported score, are all examples of errors that should not occur. To err is human, but errors should be minimized through the use of adequate quality control procedures. Practitioners should have a broad knowledge of quality control practices, as these are critical to the accurate use of tests. We believe that this document will also make an important contribution to Continuous Quality Improvement (CQI), an area in which we are striving to make progress, step by step.

The Quality Control (QC) Guidelines presented below are intended to increase the efficiency, precision and accuracy of the scoring, analysis and reporting (SAR) process of testing. They have a twofold function: They can be used on their own, as specific guidelines for quality control in scoring, test analysis and score reporting; and they can also be regarded and used as an extension of specific parts of the ITC International Guidelines for Test Use (2000). It is recommended that the reader be familiar with the ITC Guidelines and with the AERA, APA & NCME Standards (2011), as well as with other relevant international and local standards.

1.2 For whom are the QC Guidelines intended

The QC Guidelines are intended to cover large-scale testing situations, primarily where a test constitutes a measure of achievement, attainment or ability (as opposed to being an indicator of preference or some other self-report measure). As such, they are particularly applicable in large-scale educational testing and large-scale ability assessments in employment testing. Much of what follows may, however, also apply to smaller-scale assessments and to assessments involving other types of tests.

The QC Guidelines are directed at those responsible for:

- test design and development
- test administration
- test scoring
- item and test analysis (including test norming and equating)
- maintaining test security
- test interpretation
- reporting test results and providing feedback to test takers
- training and supervision of test users
- designing computer systems and programs to deal with test data
- policy making (including legislators)
- test publishing

Knowing more about quality control is essential for any professional involved in the testing process. Although intended primarily for professional use by practitioners in the field of testing, the QC Guidelines incorporate basic principles of good practice and, as such, are also relevant to those who use tests solely for research purposes in the field or in laboratory settings.

1.3 Contextual and international factors

The QC Guidelines are intended for an international audience of professionals who use testing in their work. These Guidelines can help professionals to develop specific local quality standards. Contextual factors, such as local or national laws or standards, existing regulations, or specific contracts between clients and test vendors must be taken into consideration when interpreting the QC Guidelines at the local level or determining their practical use in a given setting. For example, in some countries confidentiality with regard to the test taker's personal information is protected by law.

1.4 Errors and the need for Quality Control Guidelines

Errors that occur during the SAR process may have serious implications in any domain of measurement – psychological, educational, occupational, and attitudinal. For example, if a large number of mistakes are made in scoring a test, score meaning and score reliability will be affected – test score reliability almost certainly will be lower, which typically can also result in lower predictive validity. In some cases, an error may result in a person with pathological behavior being wrongly identified as a person with normal behavior. In other cases, errors may prevent a qualified candidate from being offered a job, or may lead to incorrect placement in academic courses. Errors may also result in misguided educational intervention, such as someone being assigned to an inappropriate educational program, or a person being granted a professional license, even though he or she lacks the required knowledge and skills. Errors can result in a huge delay in score reporting, which in turn can cause grave problems for those who, as a result, may be prevented from registering at the educational institution. In short, mistakes can have serious, even damaging outcomes. Errors can also result in a public loss of confidence in educational and psychological tests, and may reduce the credibility of certain tests if attention is given to them by the media. Errors may, in some cases, even lead to legal action being taken against the testing agency, educational institutions, or companies seeking to hire qualified employees.

Professionals who implement the testing process are subject to potential pressure from four sources: organizations, test takers, testing companies and the media. All four expect fast and inexpensive test development and rapid delivery of scores, so that they can be used as soon as possible. To maintain quality standards, it is imperative that one resist pressure brought to bear by those who seek to have the process shortened or speeded up, or to have some of the stages omitted. There is extreme pressure, for example, when an organization is bound by contract to score, analyze, and report scores within a narrow frame of time.

There is significant potential for error in lengthy production processes such as test development, scoring (especially when there is a large number of test scripts), test analysis, and score reporting, which comprise sequential stages, each heavily dependent on the previous one. Using quality standards helps prevent errors. Quality standards should be routinely monitored and updated regularly.

1.5 Scope of the QC Guidelines

The QC Guidelines focus on large-scale testing operations where multiple forms of tests are created for use on set dates. However, they may also be used for a wide variety of other testing situations (e.g., individual testing for career guidance or personal development) and assessment techniques (e.g., multiple-choice tests, performance assessments, structured and unstructured interviews, assessed group activities) and for almost any situation in which assessment occurs (e.g., for educational purposes, or at employment assessment centres). While some of the QC Guidelines are specific and relate to various individual or group-administered standardized tests, other elements in the QC Guidelines have a much wider application (e.g., in clinical, educational and occupational testing). A large number of professions engage in assessment practices (e.g., medical and rehabilitation, forensic, educational needs, employment-related), and here too the QC Guidelines can be very useful.

The QC Guidelines are applicable in any form of test administration (including paper & pencil tests and the ever-increasing computerized assessments via the Internet or offline). Test construction, test choice and test administration are not the focus of the QC Guidelines. However, the usefulness or successful application of the QC Guidelines in test scoring, test analysis, and score reporting is contingent on the test itself being appropriate and on the scores being reliable and predictive of well-defined outcomes. Allocating resources for quality control constitutes an investment in responsible practice, accountability, and fairness – most of which are major elements of any code of ethics.

1.6 Definition of quality control

For the purposes of this document, quality control can be characterized as *a formal systematic process designed to help ensure that high quality standards are maintained at all stages of scoring, test analysis, and reporting of test results, and to thereby minimize error and enhance measurement reliability.*

1.7 Examples from other professions

Quality control procedures are used in many other professions, such as engineering, aviation, software development and medicine. With regard to the field of medicine, it is interesting to note some of the contributing factors behind errors that occur in hospitals. These include the inappropriate storage of medicines, the complexity of medical interventions, new technology, flawed communication, poor team work and the lack of a clear safety policy. Such an example has an analogy in the field of testing, where the focus is the test, with all its potential for these sorts of errors to arise in the process of administration and evaluation.

1.8 Structure of the QC Guidelines

The QC Guidelines consist of two major parts:

1. General principles – general points that should be considered and agreed upon prior to scoring, test analysis, and reporting of test scores
2. Step-by-step working Guidelines

These sections are followed by two supplementary sections: a short summary, and references

2. GENERAL PRINCIPLES

2.1 Verifying quality control standards currently in use

2.4.1 Determine what quality control guidelines currently exist for tests in your organization or in your country. If necessary, formulate test-specific quality control procedures before the test is administered. Review, update and modify guidelines whenever changes in the process are made, and also from time to time as a routine check.

2.4.2 Ensure that adequate quality control procedures are in place before administering the test.

2.4.3 When dealing with a new test, consider performing a pilot simulation for the whole SAR process. Where no pilot has been performed, treat the first administration as a trial run and be ready to make improvements before subsequent test administrations.

2.4.4 Create test-specific standards for each test, in cases where they do not yet exist.

2.4.5 Create test-specific standards for each new test at the time of its construction.

2.2 Basic preparations and agreements between persons involved

Before administering the test, basic principles should be agreed upon by the professionals responsible for the test, including those responsible for test construction, administration, scoring, equating, interpretation, validation and reporting. Indeed, though they have different responsibilities and roles, the work of all the professionals involved – whether vendors, clients or partners – should be coordinated; good communication between people in different roles should enhance the quality, purpose and use of the test.

The professionals should:

2.2.1 - identify all the stakeholders in the testing process and agree who is responsible for decision making with respect to the different parts of the testing process.

2.2.2 - determine and state the purpose or purposes of test use (e.g., selection, measuring achievement, research).

2.2.3 - agree on the timetable for the SAR process.

2.2.4 - establish the best means of communication between persons or teams (where more than one team is involved), for example, the best way to convey relevant information from one team to another, or to transmit detailed descriptions (test structure, test key, etc.) from the test development team to the test analysis team.

2.2.5 - establish the best means for communicating with the client about the testing process.

2.2.6 - decide on methods for transferring assessment data to those responsible for the SAR process, for example, data obtained using an optical reader or scanner for paper & pencil tests, or electronically obtained data for computerized tests.

2.2.7 - define the weights to be used for the subtests (when they are used) and provide the rationale for choices made. One should also be ready to modify the weights after receiving the data.

2.2.8 - agree upon scoring instructions, that is, on the number of credits to be given for each correctly answered item, and decide how to deal with wrong answers. One should also be ready to modify the instructions after receiving the data.

2.2.9 - choose a scoring scale and determine the range of scale points.

2.2.10 - decide how to deal with missing data (e.g., cases where test takers have overlooked an item or mistakenly skipped a line when shading in answers, or cases where an assessor either forgets to assess a specific test taker or does so in a non-standardized manner with no possibility of repeating the assessment).

2.2.11 - define and describe the equating model, design and sample sizes needed if the scores of different test versions must be put on the same scale, as well as the equating methods used.

2.2.12 - define and describe the standard setting model, and the design and sample sizes needed if standard setting procedures are used.

2.2.13 - agree upon the degree of detail with which scores should be reported to the test takers and institutions involved, and what additional information regarding score distributions and score use should be delivered.

2.2.14 - determine which specific individuals, bodies or institutions should receive test results, ensuring compliance with legal constraints regarding data privacy

2.2.15 - determine whether reports can or should provide other personal information (e.g., whether the test content was modified, how many items were completed, what accommodations for disabilities were offered).

2.2.16 - agree upon the level of documentation needed for the whole process.

2.2.17 - agree upon the level of replication effort to be allocated to critical processes (e.g., raw-to-scale conversion tables).

2.3 Resources

2.3.1 Confirm that there are adequate resources (cost, time and personnel) available for efficient and appropriate scoring, test analysis and reporting of scores.

2.3.2 Check available backup for each resource (e.g., if the equating specialist cannot do the equating, determine who will do it instead; or if the answer-sheet scanner is out of order, locate an alternate scanner).

2.3.3 Be aware of timing problems that can occur if backups are used. Consider contingency plans to cover the unexpected absence of key personnel.

2.3.4 Allocate tasks to appropriate members of the team: Who will take care of test scoring, analysis, and reporting of test scores? Who is in charge of the whole process? The professionals in charge of testing must determine, for example, whether the individuals involved in each step of the process have the skills needed for the work they are to carry out; they must also specify requirements and specifications, and define level of process automation.

2.3.5 Determine the necessary time resources: establish a timetable for each step in the SAR process. The deadline for finalizing the process and reporting the scores should be realistic.

2.3.6 Determine the necessary software, computer and network resources: copyrighted and custom-developed software, laptops, personal computers, main frames, disk space, server space, bandwidth analysis, and so forth.

2.3.7 Determine the necessary workspace resources – is there a sufficiently large work area (with enough rooms, tables, chairs, etc.) for all staff and test takers?

2.3.8 Determine the steps needed to keep the data safe and secure electronically.

2.3.9 Ensure that any additional equipment needed (e.g., hand scoring key, calculators) is available.

2.4 Demands and expectations of stakeholders

Those who make use of the test scores – test takers, parents/tutors, teachers/counselors, as well as those managing the testing (an agency if there is one involved) – have specific demands and expectations with regard to the scoring and equating processes and the time allocated for reporting the scores. These demands and expectations should be reasonable and communicated between parties (see also the ITC International Guidelines for Test Use, 2000, Appendix B – Guidelines for developing agreements between parties involved in the testing process)

2.4.1 Where appropriate, formulate the agreement between the parties involved – stakeholders, vendors, test takers, clients, and others – in consultation with the professionals responsible for scoring, equating and reporting. Be aware that changes are made in the contract from time to time.

2.4.2 Agree upon who has final responsibility and authority to decide how to proceed when problems occur and how to resolve them.

> *For example, when a multiple-choice question does not have a correct answer, an interviewer is very arrogant, or test takers are disturbed by a noisy environment. Another case is when a question has been constructed for which only one answer was thought to be correct, but a test taker later proves that one or more other answers are also correct.*

2.4.3 Decide in advance on the process for dealing with cases where a mistake is discovered after scores have been released.

2.4.4 Provide test takers with the opportunity to question the correctness of proposed answers and to challenge their scores, or provide test takers with an opportunity to raise issues and ensure that these are addressed.

2.4.5 Have a document that can be used to defend the scoring of each item that appears on the test.

2.5 Professional staff and working atmosphere

Where a group of people are involved in the testing process it is important that they work well together. Therefore, when hiring new employees, the ability of the newly formed team to work together in harmony should be a key consideration.

2.5.1 Ensure that those responsible for test scoring, equating, and reporting are professionals who have the requisite skills and knowledge of the SAR process. Ensure that staff members have the required competencies for the job.

2.5.2 Avoid unreasonable pressure on individuals for speed of performance.

2.5.3 Avoid excessive work hours.

2.5.4 Try to cultivate a meticulous, attention-to-detail work approach (especially with regard to error prevention), but one that is also relaxed. A calm but purposeful working atmosphere is most effective in maintaining high standards.

2.5.5 Support staff by providing professional development and training, and in some cases also personal growth and social skills training, (for example, opportunities for staff to participate in system testing based on previous year data preparatory to processing current year data.)

2.6 Independent monitoring of quality control procedures

Assign to one or more professionals (depending on the size and complexity of the project) the task of monitoring adherence to QC procedure and ensuring that all issues, problems or errors are recorded. QC monitors should operate independently of those involved in the routine test scoring, analysis and score reporting processes. Monitoring should be carried out in collaboration with all stakeholders, with the aim of auditing specific processes, for example, monitoring inter-rater reliability and checking data entry error rates.

2.7 Documenting and reporting mistakes

2.7.1 All those involved in the testing process should follow agreed procedures regarding the documentation of activities and of errors or issues that arise.

2.7.2 Agree in advance which member of staff is responsible for each stage.

2.7.3 Document all activities. Use standard check sheets to show that each process has been carried out and checked off accordingly.

2.7.4 Document in detail all mistakes or errors (regardless of whether the cause is already known), beginning with the nature of the mistake, who discovered it and when, what are/were the implications, and what steps have been/will be taken to deal with it. Also document cases in which mistakes were discovered before any harm was done.

2.7.5 Advise other professionals of mistakes in an appropriate and timely manner, sometimes in a special meeting devoted to error prevention.

2.7.6 Document how to prevent future mistakes or errors.

3. STEP-BY-STEP WORKING GUIDELINES

These guidelines suggest the steps that should be taken in managing test scoring, analysis, and reporting. Where large-scale testing is undertaken, each stage should be considered and carried out with care. A pilot run of the scoring procedures should be undertaken before working with real data, so that results will be expedited. Where thousands of people are to be tested these guidelines should be explicitly followed. Where only tens of people are being tested, the principles of the guidelines should be implemented, with some stages omitted or simplified. This is because some of the procedures are resource intensive and are based upon a model with large samples. These procedures should be adapted accordingly to smaller test samples.

3.1 Design of the reporting

Before implementing the different steps, agreement should be reached with regard to reporting, which is the final product of the system. It should be decided what to report, in how much detail, to whom, when, etc. It is not sufficient to report the score to the institution or the test taker as a number or a derivative (stanines, etc.). The proper interpretation of the score is very important. Indeed, the test development, scoring and analysis stages should all take into consideration the final product – the reported interpretation of the scores. In this sense, the underlying aim, or tacit first step of the whole process of test development, is ensuring that the reported score will be properly understood. The different aspects of score interpretation should therefore be addressed from the outset. Agreement should be reached with those all involved, regarding the reporting of sub-scores in addition to a single score: Should they be reported? Will they be used?

3.2 Background data

Background and biographical data can be very useful in the quality control process for the following purposes: verifying test taker's identity, understanding unexpected results, and establishing matching groups for equating purposes. The following steps are recommended:

3.2.1 If the legal context allows for it, gather background and biographical data (age, gender, ethnicity, education, scores on previous tests, etc.) by means of advance registration, on-site registration, or after the exam has taken place, by approaching test taker or institution.

3.2.2 If possible, check the biographical data of the test takers periodically and systematically; pay attention to inconsistencies for test takers who are retested.

3.2.3 Conduct studies to determine the expected correlation between background data and scores, look for inconsistencies in the patterns of scores in the current data with respect to other information – previous data sets, research findings etc. For example, it could be that adults perform better than young people on a certain test. If a study had indicated that young people would be expected to perform better on such tests, then the scoring process should be examined to see if a mistake has occurred.

3.3 Scoring

3.3.1 Obtaining and storing of all test takers' answers

Where relevant, all hard-copy answers of the test takers should be kept, and also stored electronically, if appropriate, usually with the identification number of each test taker. Storage of such materials – in print and electronically – will be for a set minimum and maximum period of time, in keeping with professional practice and local legal requirements. This applies to personally identifiable answer sheets and electronic records of test responses or scores and records of score information, however derived and however stored.

3.3.1.1 If paper & pencil copies exist, they should be saved for a time period in accordance with the laws of the country, state, or province in question, where such laws exist.

3.3.1.2 With regard to electronic versions, use the uninterruptible power system (UPS), and backup batteries for the computers or other technological means to reduce the likelihood of "interruption" problems and loss of data.

3.3.1.3 Scanners, where used, should be regularly checked and calibrated.

3.3.1.4 Conduct routine manual checks on scanner output.

3.3.1.5 Check the test takers' database to maintain a rigorous ID system. For example, look for cases where an ID number appears for more than one name.

3.3.1.6 All data should be protected and safe. Protect personal information by separating identifying information (e.g., names) from scores where possible. For example, keep separate files: one with biographical data, and one with scores which can be merged by ID. All these actions should comply with national laws regarding data security and data storage.

3.3.1.7 Perform check of accuracy of scoring algorithms and proper loading/use of conversion tables and norms.

3.3.2 Scoring

Once the test data have been processed and safely stored in a data base, test takers' answers are usually used to compute raw scores. In Classical Test Theory (CTT), for example, the raw score usually equals the number of answers which are correct when there are right/wrong answers; sometimes a correction for guessing is applied, sometimes there are items that weigh more than the others. In Item Response Theory (IRT) the raw score is the latent ability – often referred to as "theta" or "trait score." Scoring can be adversely affected by many different kinds of errors, such as those resulting from an incorrect key. Sometimes errors lead to extremely low scores. Use the following quality control procedures to catch such errors:

3.3.2.1 Check whether the data structure complies with the requirements in data record layouts, e.g., order of items in the file.

3.3.2.2 Apply approved rules to remove invalid cases, recode missing information, and handle duplicated cases.

3.3.2.3 Compare sample data to the range of scores that can be expected, and compare descriptive statistics to the test publishers' norms, if those are provided. Sample statistics can be expected to deviate somewhat (beyond what is expected by sampling error variance), but large effect size differences should be noted and potentially investigated.

3.3.2.4 Review extreme scores of individuals or defined groups – both low and high – (for paper & pencil and computerized tests). Extreme scores may indicate three possible problems: a mistake in score computing, dishonest acts, or a mistake in obtaining the data.

3.3.2.5 Review the data of individual test takers in cases where the variation between correlated subtest scores is larger than expected.

3.3.2.6 Run an item analysis and examine item statistics – errors in the scoring key for a single item are hard to detect, unless item statistics are available (mis-keyed items often appear difficult and might appear to be discriminating negatively, e.g., may appear to have negative correlation with a criterion).

3.3.2.7 Check omit rates for every item. An item may have been removed by mistake from scoring for some of the examinees.

3.3.2.8 Pay special attention to groups who had different test conditions, and perform additional checks on this data. For example, people who were tested on a different date, or with a different test version, or used a different response method.

3.3.2.9 Compute and review basic statistics for major test taker units, for example, by test hall, test administrator, or by computers using the same Internet connection. An error may result from a wrong test form having been assigned to a specific test hall.

3.3.2.10 If resources permit, a random sample of answers sheets should be given to a different team to analyze and score. Later, a comparison between team results can be made.

3.3.3 Rating performance tests, work samples, role plays, interviews, etc.

Whereas the scoring of multiple-choice (MC) items is objective (based on a defined key) and highly reliable, the scoring of open-ended (OE) items (performance assessment, OE questionnaires, work samples, role playing etc.) usually has a subjective component. This kind of scoring thus tends to be less reliable than MC scoring, because OE often involves human raters and is affected by their input. Nevertheless, a variety of means can be used to reduce the subjectivity inherent in the scoring of OE items, as well as to improve the reliability and accuracy of OE scoring.

3.3.3.1 Make sure that performance on tests, work samples, role playing, and interviews are rated by trained assessors who have the requisite knowledge and experience, as well as credentials, training or appropriate formal education.

3.4.2.1 The instructions for rating OE responses should be clear and well-structured.

3.4.2.2 Use range-finding activities to pinpoint examples of OE student's responses at each rubric point. Involve sample papers in scoring training activities.

3.4.2.3 Require raters to participate in training sessions before undertaking the rating process. Training enables them to become familiar with rating instructions and to practice the scoring of assessment material before they become authorized to evaluate actual test taker responses.

3.4.2.4 Assess raters' competence based on their training, prior to having them undertake operational rating.

3.4.2.5 Try to use at least two raters for each assessment, depending on costs and availability of resources.

3.4.2.6 When there is only one rater for all test takers (due to financial or other considerations) use two raters per sample (e.g., 10% of the data) to estimate scoring reliability, depending on stakes, length of test, and other factors.

3.4.2.7 If computer scoring of OE items is used, ensure that the scoring is monitored by a human rater. Justify the use of computerized scoring on the basis of research before using it operationally.

3.4.2.8 Ensure that raters work independently of one another.

3.4.2.9 Apply statistical procedures to assess the reliability of the rating process, i.e., by computing appropriate measures of inter-rater agreement as well as differences between raters within and across raters by checking the degree of correspondence as well as the differences between raters and using appropriate measures to eliminate correlation coefficients between rater results that are similar just by chance.

3.4.2.10 Monitor the rating quality periodically in real time, so feedback will be available.

3.4.2.11 If a rater is not meeting expectations, (ratings are unreliable or not close enough to those of other raters) inform the person accordingly and consider retraining; do not hesitate to replace the person if the problem is not resolved.

3.4.2.12 Develop policies for dealing with large discrepancies between raters. When differences are small, they should be averaged or summed to avoid rounding problems. When there are large discrepancies, an experienced rater may mediate to resolve them.

3.4 Test Analysis

3.4.1 Item analysis, usually for large-scale multiple-choice and open-ended testing

Item analysis provides basic statistics for making decisions about the characteristics of the items and how they perform when grouped together to form a total score. It is recommended that item analysis be performed on every occasion and for every test form, unless the number of test takers is small. The item analysis statistics consist of item difficulty (or item 'acquiescence' – with personality questionnaire items) and discrimination. IRT parameters for each item can be calculated in many instances, depending on the model used for the development of the test. In addition, item analysis presents general test statistics (reliability and/or standard error, mean, standard deviation, test information, distribution of test takers' responses, etc.). The following procedures should be observed whenever the number of individuals taking the test exceeds a minimal number, depending on the model used.

3.4.1.1 Use a reliable process to perform the item analysis, and make sure that the programs have adequate technical documentation.

3.4.1.2 If you have reason to believe that the item analysis program is fine or you are using a new program, use two programs simultaneously and compare the results.

3.4.1.3 Conduct item analysis after the exam is administered or analyze accumulated data if the exam is given periodically. Consider performing item analysis on partial data, (before full data is available), so you can identify errors quickly.

3.4.1.4 Review item analysis before drawing conclusions about test takers.

3.4.1.5 Item analysis will identify many answer-key problems in a test. For instance, what seems like a very popular distractor may in fact be the correct answer (key); or negative inter-item correlations may indicate an item that was not reverse-coded. If the results for a particular item are problematic, the key and content of that item should be reviewed.

3.4.1.6 Repeat item analysis if the test key is modified or if items are deleted. Update documentation (e.g., scoring tables and equating specifications) throughout the process.

3.4.2 Equating/Calibrating new test forms and items

Sometimes, equating is not important because applicants only compete with those tested at the same time using the same test version. Without equating, applicants' scores from previous test administrations cannot be carried forward and compared with new form test results for other applicants. If scores for different test versions need to be on the same scale, new test forms should be equated so that the test scores achieved on these newer forms will be comparable to old test forms that may have different psychometric characteristics. The outcome of the equating is that scores from all the test-forms have the same meaning. Equating can be done before the test is administered (pre-equating), and/or after test administration (post-equating). Equating can be conducted using data at the item-level, scale-level or test-level. There are different perspectives and methods for equating (e.g., linear equating, equipercentile equating, IRT-based common-person or common-item equating).

Equating usually requires large samples, depends on the equating method and design (See Kolen & Brennan, 2004; Lamprianou, 2007).

3.4.2.1 In cases of unexplained equating problems (e.g., scores are lower than expected), confirm that all test forms were administered under the same standardized conditions. If the conditions were not standardized, try to estimate the impact of the different conditions.

3.4.2.2 Develop procedures to ensure that the specified equating procedures and designs have been followed correctly.

3.4.2.3 Explore the assumptions of the equating procedure and/or determine whether different equating procedures that have different assumptions yield similar results. Perform a common item parameters stability check. If you are using a common item set for equating, document the rationale followed when dropping some of the common items from the common item set, and the effects of such decisions on scores and on location of equated cut scores. Also document content representation and statistical characteristics of the common item set after the screening. This standard also applies to the common person design; but the screening is focused on test takers.

3.4.2.4 Compare the scores obtained for the test takers with those that were anticipated on the basis of the test taker's background (see 3.2.1). If there are discrepancies recheck the scores.

3.4.2.5 Perform historical comparisons of scores and passing rates. For well-developed large-scale assessments, fluctuations between year-to-year administrations are often small. Large rates may signal a problem in the equating of the test scores, a change in population characteristics, for example.

3.4.2.6 When there are multiple test administrations (many administrations with a small number of examinees in each in contrast to few administrations with a large number of examinees) apply specific quality control tools for monitoring the stability of test scores. Some of the tools are: Shewhart and CUSUM charts on testing data, time series models, change point models and data mining tools (see Von Davier, 2011).

3.4.2.7 If there are cut scores that divide the test takers on a pass/fail basis or other performance levelling basis, check the pass-fail and other level rates. Compare them to the anticipated rates based on previous years, test taker background and similar tests.

3.4.2.8 Ensure consistency across panels in setting cut scores, use defensible methods and document the process. Also document cases where you deviate from the standard process.

3.4.2.9 If a different test administration format is used (e.g., computerized administration of a paper & pencil test) it is necessary to compare the new test characteristics to the old, and sometimes to equate the new test to the old.

3.4.2.10 For high-stake tests make every effort to replicate equating results independently and involve a third party external to the equating process.

3.4.3 Computing standardized scores

In many instances, standardized scores are used to help make scores comprehensible. In such cases, raw scores from the test are used to compute whichever standardized scores are to be employed (e.g., stanines, deciles). Parameters or conversion tables are used to compute the scale and the standardized or percentile score to be reported. Usually raw scores (number-correct or number-correct adjusted for guessing) or theta scores (IRT-based tests) must be converted to the specific test scale. The conversion is done by means of a lookup table or a function (e.g., for linear transformation).

3.4.3.1 Perform appropriate raw score conversions to obtain the defined score scale

3.4.3.2 Check score scale-conversions and other clerical procedures for accuracy.

3.4.3.3 Check that the correct conversion has been used.

3.4.3.4 Verify that low standardized scores are based on low raw scores and that high standardized scores are based on high raw scores.

3.4.3.5 In some cases additional procedures should take place at a later point, following the conversion (e.g., defining a uniform minimum and maximum in each reported score.)

3.4.3.6 Compare the properties of the new test form with other test form tables/parameters for unusual discrepancies or for similarity.

3.4.3.7 Account for changes that occur in the scale over time.

3.4.3.8 Calculate some scores manually and compare results with the computer-generated results.

3.4.3.9 Check the statistical relation between raw scores and standardized scores, using a scatter plot.

3.4.3.10 Use two different programs to calculate standardized scores and compare them.

3.4.3.11 In the technical manual or associated materials, give a detailed description of the procedures used to convert raw scores to standardized scores. Because this technique may be different for different test forms, the procedure should be described for each test form.

3.4.4 Security checks for tests

If a score is reported and subsequently discovered to have been obtained by cheating, it creates a serious problem that compromises the security and integrity of the test and the testing system. Unfortunately, cheating cannot be completely prevented, even with monitoring and other deterrents in place; the temptation to cheat can be immense, especially in high-stakes testing. As part of the continuous battle against cheating, lawyers should be consulted to review the security checks and confirm their applicability. In high-stakes national educational testing, fraud can be perpetrated at the individual level, but also at the level of the class, school, district or workplace. It may occur at the testing site, via cell phones, or though websites on the Internet. In the employment domain, as testing for a job from the candidate's home (via the Internet) becomes more widely used, the risk of impersonation and various forms of dissimulation increases. A side benefit of security checks – apart from identifying cases of fraud – is that they are also sometimes an indicator for problems in test administration, data collection or storage. The following precautions are suggested:

3.4.4.1 Verify the identity of test takers as they are admitted to the exam hall or when there is testing at home, use a picture identity card (ID) or biometric means such as an iris scan or fingerprints. Use advance techniques to identify distance test takers (see Segall, 2001ASVAB)

3.4.4.2 It is better to use multiple test forms. When one test form is used, do not seat test takers who may be acquainted (e.g., live in the same neighbourhood, attend the same school) close to one another. For example, seat test takers by alphabetical order.

3.4.4.3 Record seat numbers to help in the copying analysis.

3.4.4.4 Where appropriate (e.g., when copying is suspected), employ statistical copying indices, based on the similarity of the answer sheet of each test taker to those of other test takers in the same test hall and in the whole testing site.

3.4.4.5 Employ trained, reliable proctors and supervise them routinely. Ensure that proctors do not have a conflict of interests.

3.4.4.6 Check aberrant or unexpected response patterns, (e.g., when difficult items are answered correctly and easy items incorrectly).

3.4.4.7 Obtain a handwriting sample from each test taker prior to and during the examination, to help in identifying impersonators or those with suspicious behaviour. This procedure may be skipped if there are no identification problems.

3.4.4.8 Analyze divergent scores on a repeat test administration (if repeaters can be located), using a statistical distribution of reasonable differences between the current score and the score that was obtained in a previous administration. Extreme differences may indicate that an impersonator took the place of the intended test taker, or that the candidate obtained information about test items in advance of the test administration.

3.4.4.9 Have a documented routine (a legal one, if needed) for dealing with suspected cheaters. Inform all test takers in advance that policies to combat fraud and cheating are implemented, monitored and enforced.

3.4.4.10 Sometimes, teachers have an interest in raising students' scores in standardized tests. For this reason teachers should not be given access to scores of standardized tests.

3.4.4.11 Use locked cabinets and secured servers to secure test materials and test results. Ensure that all those in involved in developing test items and tests are trustworthy and fully briefed regarding security requirements.

3.4.4.12 Computers that are used for test presentation need to be locked down to disable the saving and sending of test materials. Connection to the Internet should be avoided if it can enable the sending of test materials.

3.5 Reporting

3.5.1 Reporting test scores

Scores are released to both test takers and test users (the client). Ideally, score reports should be provided in printable form. In some instances and settings, the Internet is increasingly being used and is becoming the standard reporting method. Reporting must be done in such a way that the meaning of the scores is clear to the test taker and to the client.

3.5.1.1 Use focus groups of test takers or possibly "think-aloud procedures," "experimental studies," or even "one-on-one interviews" to gain information to assist in the development of comprehensible and instructive explanations of the score report and any interpretive guide.

3.5.1.2 Ensure that anyone who receives the scores has appropriate guidance in interpreting them, so there will be a proper understanding of test scores. Support this with evidence that the reports allow users to make defensible interpretations.

3.5.1.3 Create computer-generated reports that can help mediate technical issues and that will be appropriate for their recipients.

3.5.1.4 Use roll-up, on-the-fly reporting data warehouses for transnational, national and province-level test results, where needed.

3.5.1.5 Clarify to what level different scores can be relied upon (e.g., where sub-scores have too low reliability to use in making high-stakes decisions)

3.5.1.6 Use the help of public relations experts when results and scores are reported to the media and politicians.

Take steps to maintain the security of score reporting

3.5.1.7 Take steps to ensure that the individual score report cannot be forged by the examinee.

3.5.1.8 Avoid editing the institution report: editing may cause serious problems. If there is a need to change one or more scores, use allocated software, or create the report again.

3.5.1.9 Encrypt the electronic files of score reports for storage and transfer.

3.5.1.10 Ensure that score reports are only sent to appropriate individuals. Do not send score reports that are more inclusive than necessary. It may be easier to send the same complete report to all test users, but to maintain candidate confidentiality, only relevant results should be sent to each test user.

3.5.1.11 Inform institutions that only the report sent directly to the institution – and not the test taker's copy of the report (which can be faked) – is to be used for official purposes. Also recommend the institutions to do routine verifications on the institution report.

3.5.2 Documentation

Routine documentation of the entire scoring process, including key descriptive statistics (mean, standard deviation, median, score range, reliability…) and comparison of current test taker groups with previous test taker groups should be conducted and completed before or shortly after the release of test scores. A good documentation culture may contribute to future processes being more reliable and accurate. Allowing the public access to information regarding the scores serves as an additional control on the whole SAR process. It is important to:

3.5.2.1 - document the entire process, step by step (internal report), with routine documentation of the entire scoring process, including main statistics and group comparisons.

3.5.2.2 - ensure that new test forms are administered only after documentation for an old test version has been completed.

3.5.2.3 - compile descriptive statistics, for example, with regard to gender differences and yearly distribution, and allow the general public access to these statistics. Short explanations of these statistics should be provided. Aggregated statistics protect the anonymity of the individual test taker.

4. SUMMARY AND FINAL COMMENTS

In addition to all the recommendations made in this document, some general guidelines and suggestions are in order. Each time a new test is introduced, a detailed real-world simulation of the entire process, step by step, should be carried out (see Texas Education Agency et al., 2004); the new testing procedures can then be practiced and evaluated. Such a simulation will inform any subsequent revised standards for quality control procedures. Furthermore, the process of scoring, test analysis and score reporting is one that consists of sequential stages, in which every stage is based on the successful completion of the preceding stage. Therefore, it is recommended that a quality control checklist based on the QC Guidelines be codified in such a way as to make it impossible to proceed to another stage without the preceding one being completed. Computerization seems like the most logical approach to standardize, modify, and control QC procedures easily, transparently and effectively. However, although the benefits of computerization are widely accepted, a trained human with research background is necessary to devise the quality control procedures and to adapt and evaluate them.

5. REFERENCES

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

Allalouf, A. (2007). Quality Control Procedures in the Scoring, Equating, and Reporting of Test Scores. *Educational Measurement: Issues and Practice*, 26: 36-43.

Bartram, D., Hambleton, R.K. (Eds.) (2006) .*Computer-Based Testing and the Internet.* West Sussex: John Wiley & Sons.

Cizek, G. J. (1999). Cheating on tests: How to do it, detect it, and prevent it. Mahwah, NJ: Lawrence Erlbaum.

ITC (2001). International Guidelines on Test Use. *International Journal of Testing*, 1: 95-114.

ITC (2006). International Guidelines on computer-based and Internet-delivered testing. *International Journal of Testing,* 6: 143-172.

Kolen, M. J., and Brennan, R. L. (2004). *Test equating, linking and scaling: Methods and practices*. New York: Springer.

Lamprianou, I. (2007). Comparability methods and public distrust: an international perspective. In Newton, P., Baird J., Goldstein, H., Patric, H., & Tymms, P. (Eds) *Techniques for monitoring the comparability of examination standards.* Qualifications and Curriculum Authority, London.

Nichols, S. L. & Berliner, D. C. (2005). The inevitable corruption of indicators and educators through high-stakes testing*, Educational Policy Studies Laboratory*, College of Education, Arizona State University.

Rhoades, K., & Madaus, G. (2003). *Errors in standardized tests: A systemic problem*. (NBETPP Monograph). Boston, MA: Boston College, Lynch School of Education.

Texas Education Agency, Pearson Educational Measurement, Harcourt Educational Measurement & Beta, Inc. (2004) *Chapter 9: Quality control procedures. Texas Student Assessment Program.* Technical Digest (2003-2004)

Toch, T. (2006). *Margins of error: The testing industry in the No Child Left Behind era.* Washington: Education Sector Report.

Von Davier, A. (2011) *Statistical Models for Test Equating, Scaling, and Linking*. Springer

Wild, C. L., & Rawasmany, R. (Eds.) (2007). *Improving testing: Applying process tools and techniques to assure quality*. Mahwah, NJ: Erlbaum.

Zapf, D. & Reason, J. (1994). Introduction: Human Errors and Error Handling. *Applied Psychology: An International Review,* 43*:* 427-432.