An NCME Instructional Module on

# Quality Control Procedures in the Scoring, Equating, and Reporting of Test Scores

Avi Allalouf, *National Institute for Testing and Evaluation*

*There is significant potential for error in long production processes that consist of sequential stages, each of which is heavily dependent on the previous stage, such as the SER (Scoring, Equating, and Reporting) process. Quality control procedures are required in order to monitor this process and to reduce the number of mistakes to a minimum. In the context of this module, quality control is a formal systematic process designed to ensure that expected quality standards are achieved during scoring, equating, and reporting of test scores. The module divides the SER process into 11 steps. For each step, possible mistakes that might occur are listed, followed by examples and quality control procedures for avoiding, detecting, or dealing with these mistakes. Most of the listed quality control procedures are also relevant for Internet-delivered and scored testing. Lessons from other industries are also discussed. The motto of this module is: There is a reason for every mistake. If you can identify the mistake, you can identify the reason it happened and prevent it from recurring.*

**Keywords:** scoring, equating, score reporting, quality control, standards

There is significant potential for error in long production processes that consist of sequential stages, each of which is heavily dependent on the previous stage. Such processes can be critically affected by variations in material, weight, timing, temperature, or other parameters, whether the task at hand is the design and manufacture of a motorcar or the baking of a cake. Making sound choices among alternatives can also be critical to the success of the entire process.

Mistakes, such as the computing or reporting of an incorrect score (one that is lower or higher than the correct score), may have serious implications in the context of educational measurement. They might preclude a qualified candidate from being accepted to college, lead to incorrect course placement, result in misguided educational intervention, or in the granting of a professional license to a person who lacks the required qualifications. Moreover, mistakes that cause real damage of this kind can precipitate legal action against the testing company or the educational institution. Finally, a high incidence of such mistakes will have an adverse impact on test reliability and validity.

*Avi Allalouf is Director of Scoring & Equating at the National Institute for Testing and Evaluation (NITE), PO Box 26015, Jerusalem 91260, Israel (www.nite.org.il). His specializations are test equating and differential item functioning.*

*Series Information*

ITEMS is a series of units designed to facilitate instruction in educational measurement. These units are published by the National Council on Measurement in Education. This module may be photocopied without permission if reproduced in its entirety and used for instructional purposes. Dr. Deborah Harris has served as editor for this module. Information regarding the development of new ITEMS modules should be addressed to: Dr. Mark Gierl, Canada Research Chair in Educational Measurement and Director, Centre for Research in Applied Measurement and Evaluation, Department of Educational Psychology, 6-110 Education North, University of Alberta, Edmonton, Alberta, Canada T6G 2G5.

## Responsibility and Accountability in the Scoring-Equating-Reporting Context

Any individual, team, company, or institute that develops, administers, scores, and reports test results is responsible for maintaining professional standards. All test programs should conform to these professional standards.

Accountability, in this context, means that the individual, team, company, or institute is responsible for developing

and implementing procedures to justify and defend all the processes carried out in the course of the scoring-equating-reporting continuum. This means, operationally, that each professional effort is made to ensure that each examinee's score, identification, test data, the test key, equating, scoring, reporting, and documenting are accurate.

### The SER Process

Educational measurement through testing involves five main elements: test development, test administration, test analysis, score reporting, and validation. This module deals with two of these elements: test analysis and score reporting. Test analysis is divided into three sub-elements: scoring, item analysis, and equating. Throughout the module we will refer to the process of scoring, equating, and reporting test scores as the SER process.

### Why Mistakes Happen

To err is human. Errors can occur in various ways, and there are several existing strategies for error prevention (see Zapf & Reason, 1994). A wide variety of factors contribute to the incidence of mistakes in the SER process. The most common are related to the management of professional standards. Such standards may be a legal prerequisite. Some of the principal reasons that mistakes occur are as follows:

1. **No standards have been established.** This is a scenario that some might consider unlikely, even unbelievable. In such cases, the person in charge of the SER process would act on the basis of his or her own judgment.
2. **Practices are undocumented** (as opposed to documented standards). The process is based on undocumented practices that vary from test to test or from one administration to another.
3. **Standards are not followed.** Detailed standards exist for every step in the SER process, but they are not maintained. There may be several reasons for the failure to maintain standards: lack of expertise (the person in charge of following the standards is incapable of doing so), lack of awareness regarding possible problems, and over-confidence (the person in charge is overly confident that "everything is fine").
4. **Standards are not updated.** This is one of the main reasons for presenting this module. Standards must be routinely monitored. They should be updated on the basis of professional knowledge and grounded in experience (which usually takes the form of mistakes. . .).
5. **Scheduling and financial pressures.** Professionals who implement the SER process are subject to tremendous pressure from three sources: educational institutions, examinees, and testing company management. The institutions want the scores immediately so that they can be put to use quickly (e.g., to complete the admissions process and notify accepted candidates before other institutions do so). The examinees are in a hurry to know their scores (which usually indicate their chances of being accepted, accredited, or graduating). The testing company management is interested in pleasing both the institutions and the examinees, and is eager to finish dealing with one test and begin working on the next.[1] Budget considerations also play an important role.
6. **Responsibility problems.** Who is responsible? Sometimes it is not clear who is responsible for each step of the process. One classic example would be a case in which there is more than one involved party, such as author and vendor.

### Are Mistakes Unavoidable?

Unfortunately, it seems that the scoring, equating, and reporting of test scores can never be entirely free of mistakes. Errors can, and do, happen during the SER process. There are so many steps involved, so many points at which mistakes can occur, that a low rate of error can only be achieved by maintaining extremely high standards at all times and applying adequate quality control procedures. This module suggests effective ways for the prevention and detection of mistakes.

### Quality Control—Definition and Components

"Quality control" has many definitions. W.E. Deming, one of the founders of the philosophy and application of statistical control of quality, defines it as follows: "Inspection with the aim of finding the bad ones and throwing them out is too late, ineffective, and costly. Quality comes not from inspection but from improvement of the process."

A relevant definition for the present purposes is as follows: *Quality control is a formal systematic process designed to ensure that expected quality standards are achieved during scoring, equating, and reporting of test scores.* This definition in hand, we still have to define the "ingredients"—the expected standards and the detailed processes—of quality control.

Kolen and Brennan (2004, p. 309) list six quality controls with which to monitor equating:

1. Check that the administration conditions are followed properly.
2. Check that the answer key is correctly specified.
3. Check that the items appear as intended.
4. Check that the equating procedures specified are followed correctly.
5. Check that the score distribution and score statistics are consistent with those observed in the past.
6. Check that the correct conversion table or equation is used with the operational scoring.

Although the above is a partial list that deals mainly with the equating process, it constitutes an excellent starting point. Reading between the lines might give some indication of the many possible mistakes that these checks can reveal. Our module goes further and provides many explicit examples of mistakes that can occur.

### Module Structure

The module divides the SER process into 11 steps. For each step, possible mistakes that might occur are listed, followed by some examples and suggested quality control procedures for avoiding, detecting, or dealing with these mistakes. Finally, other recommendations—such as application of simulations and management of human and computer resources—are discussed. The module does not address the test construction process, which also holds potential for errors, and assumes that the relevant quality control procedures were properly applied.[2]

*The Eleven Steps of Scoring, Equating, and Reporting*

First, we will list and explain the 11 steps. The steps are ordered chronologically; however, they may overlap or occur in a different order in certain tests. Most of the steps are relevant for both paper and pencil and computerized or Internet-delivered tests.

1. *Knowing your examinees in advance*—Knowing the examinees' background before the test administration.

**Scoring**

2. *Obtaining the examinees' answers*—Optical reading or scanning of the answer sheets for paper and pencil tests, saving of keyboard records for computerized tests.
3. *Storing examinee data in a database*—All data should be stored in the database according to examinee name, identity number, the exam room in which the testing took place, and/or other variables.
4. *Scoring* (temporary raw scores)— Raw data are translated into raw scores.
5. *Item analysis*—Prior to equating the test, an item analysis should be conducted. This provides the distribution of responses and the relationship between the responses and a criterion of interest for each item.
6. *Computing final raw scores*—After problematic items are removed from the equating and scoring procedures (based mainly on the item analysis), final raw scores are computed.

**Equating**

7. *Equating* new *test forms and items* (*old* test forms and items are usually not equated)— There are two kinds of equating: pre-equating, and post-equating. In addition, equating can be done using item-level and/or test-level data.
8. *Computing standardized scores*—Using parameters and conversion tables to compute the scale, percentile, or standardized score to be reported.

**Reporting**

9. *Test security checks*—"One minute before reporting"—Detection of fraud (such as copying and impersonation). Suspicious examinees are removed from score reporting until their performance is verified.
10. *Reporting test scores*—Delivering the scores to the examinees and to the client institutions electronically and/or in printed form.
11. *Documentation*—Routine documentation of the entire scoring process, including main statistics and trends.

**Elaboration on the Eleven Steps**

1. **Knowing your examinees in advance**—Knowing the examinees' background before the test takes place (this can be accomplished only when there is pre-registration for the test).

*Explanation*
It is important to know the basic demographic characteristics of future examinees, as there is often an association between examinees' background data and their scores. Gathering these data can be accomplished only through advance registration for the test (as opposed to a "walk in" procedure). Basic characteristics include age, gender, and cognitive characteristics such as scores on previous tests, and educational background. It should be noted that background data are used solely for explaining unexpected results.

*Mistakes or indications of the need for quality control—[Examples]*
Demographic data may help solve some of the following:

1. After a test is administered and scored, it is found that the score profile is unusual: scores at a specific area are found to be much higher than the scores at another specific area.
2. There are problems in equating, probably due to the fact that the two populations that served for equating differ greatly in their ability.

*Quality control processes*
1. During registration for the test, examinees should routinely be asked to report their background data.
2. Another possibility, if there is no pre-registration, is to ask the examinees to report their background data during the walk-in procedure. Usually, in this situation less data will be gathered.
3. Look for statistical relations between background data and scores. One of the routine quality control procedures is comparing the obtained scores with anticipated scores.
4. Try to explain any irregularities by means of the relation found; irregularities can be explained by examinee background; for example, age and previous scores on various tests.
5. The examinees' background can be used to form a matched ability group for equating purposes. A matched group can be useful if the original equating led to inaccurate equating results (in cases with two groups of differing ability).

2. **Obtaining the examinees' answers**—Having all the answers in an electronic medium

*Explanation*
All of the examinees' answers should be saved. Hard copies, if they exist, should be saved for several years (in accordance with legal requirements). In addition, answers should be stored electronically—optical reading or scanning of the answer sheets for paper & pencil tests, saving of keyboard records for computerized tests.

*Mistakes or indications of the need for quality control—[Examples]*
1. An answer sheet is lost.
2. Data from a computerized test are lost because of a power outage.
3. An examinee accidentally records the answers on an answer sheet that belongs to another (absent) examinee.

Optical mark readers (OMR) and optical scanners cannot be considered perfect; nor can OCR (Optical Character Recognition) technology.

*OMR-specific problems*:
1. Faint marks might be treated as missing.
2. An erased answer might be treated as a legitimate mark.
3. When there appear to be two marks on the same item, the scanner might not differentiate between the intended mark and the erased one.

*OCR-specific problems:*
1. A character is wrongly identified as another (similar) character.
2. The ID number is not recognized.

*Quality control processes*

1. Use UPS and backup batteries for the computers.
2. Machine maintenance—check the scanner and calibrate it periodically.
3. Have backups (another machine and/or a manual alternative).
4. Make sure that all answer sheets have been processed by comparing the number of answer sheets to the number of examinees present.
5. Use smart software to review the scanner output and achieve maximum accuracy.
6. Check faint marks manually.
7. Review low scores manually.

3. **Storing the examinee data**—Data are stored in a database

*Explanation*

Data are usually stored according to examinee name and identity number. Information regarding gender, birth date, other demographic data, exam date, payment confirmation, exam type, and exam version is also usually stored. Data regarding the test administration (test hall, date, testers' ID, and number of examinees per class) should be stored as well.

*Mistakes or indications of the need for quality control—[Examples]*

1. Two different examinees (with different names) have the same identity number. Sometimes one of the examinees cannot be located to confirm his identity.
   1.1 This may be due to incorrect registration of one of the examinees.
   1.2 This may be caused by a mistake in the scanning of the answer sheet of one of the examinees.
2. An examinee's identity number is incorrect.
   2.1 Shifting—The number was entered with a shift of one position to the left or right.
   2.2 Typo—The number was entered with one or more incorrect characters.
3. The examinee's test data are obtained from different sources and cannot be matched: you have an examinee's data from one source (e.g., one test) but you do not have it from the other source (e.g. another test).

*Quality control processes*

Keep the database accurate (routinely) by means of the following:

1. If there is a connection between the ID and certain demographic data, the ID may be validated by means of these data.
2. Perform routine checks on the examinee database for cases with different names but the same ID (additional information: gender, date of birth, place of birth).
3. If necessary, contact examinees to verify their ID numbers.
4. Validate every ID against official, state, or national ID databases, where possible.
5. Use an embedded ID check in the answer sheet.
6. Keep detailed logs of the scoring process. In case of a mistake, this will enable analysis of the cause of the mistake.
7. Create an intelligent merging system if the data come from different sources.

4. **Scoring** (temporary raw scores)—Raw data are translated to raw scores

*Explanation*

The examinees' answers to the dichotomous or polytomous scored items are used in the computation of raw scores. Sometimes, a correction for guessing is applied. These scores will be used later on (Step 8) for computing the standardized scores.

*Mistakes or indications of the need for quality control—[Examples]*

1. A wrong answer key (for an individual or for a group of examinees).
   1.1. A wrong answer key for one of the items
   1.2. A wrong answer key for an entire test
   1.3. An item has no correct answer
   1.4. An item has two correct answers (contrary to what the examinees were told)
2. The wrong test form is given to an examinee.
3. The number of items in the test was increased but the test key was not updated.
4. Pages in the test booklet not in order.

*Quality control processes*

1. Check examinees with low scores (also check the optical reading).
2. Check examinees with large differences between scores on different parts of the test.
3. Perform checks for special groups (e.g., examinees who arrived late for the test).
4. Perform checks on a random sample.
5. Compute and review basic statistics on major units such as examinees test halls.
6. Perform item analysis (see next step).

5. **Item analysis**—After test administration, an item analysis is performed

*Explanation*

Prior to equating the test, item analysis is performed to provide basic statistics (difficulty, discrimination etc.) for each item. This is in addition to test statistics (reliability, mean, standard deviation etc.). The item analysis presents some or all of the following statistics and parameters: the distribution of examinees' responses, the relationship between the responses and a criterion of interest, and IRT parameters.

*Mistakes or indications of the need for quality control—[Examples]*

Skipping item analysis is a critical mistake. It may happen in the following situations:

1. An "old" test form is administered. The psychometrician is (incorrectly) assured that there is no need to perform item analysis, since the test key has been checked before. However, a typo has occurred in the test booklet, causing an unintentional change in the test key.
2. Item analysis is performed but the analysis output is not reviewed.

*Quality control processes*

It should be noted that the item analysis itself, whether based on IRT parameters or CTT statistics, should be reviewed before conclusions are drawn. A possible review would be the comparison of CTT statistics to IRT parameters (e.g., correlating CTT difficulty indices to IRT b parameters; a high correlation is expected).

Item analysis is capable of detecting most answer-key problems in a test. It is the best tool for detecting problems such as an incorrect key, two correct answers, or an item with no correct answer. If the analysis of an item is problematic, its content should be reviewed. Item analysis must never be skipped, even when previously used test forms are administered.

When the sample is small, and regular item analysis is unsuitable, a modified version of item analysis should be used.

6. **Computing of raw scores**—(to be used for the final, standardized, scores)

*Explanation*
After problematic items are removed from the equating and scoring procedures (mainly as a result of item analysis), and after all the other controls have been completed, final raw scores are computed.

*Mistakes or indications of the need for quality control*
Problematic items that ought to be removed from the equating and scoring procedures (based on the item analysis) are not removed.

*Quality control processes*

1. Repeat the item analysis (if there were changes in the key).
2. Check that problematic items have indeed been removed.
3. Perform a similar check to step 4 (low scores, differences between domains, special group checks).

7. **Equating of new test forms and items**—So that they are on the same scale as old forms and items

*Explanation*
There are two kinds of equating: pre-equating, which is done before the test is administered, and post-equating, which is done after test administration. Equating can be conducted using item-level or test-level data.

*Mistakes or indications of the need for quality control*
Equating rests on the assumption that the test administration conditions were standardized. It also relies on statistical assumptions that are usually only partially affirmed; for example, the assumption that the common items used to equate between test forms have the same characteristics in each test form. This situation, in which statistical assumptions are only partially met, led equating specialists Kolen and Brennan (2004) to state that controlling of equating is critical and must always take place. Sometimes, a choice must be made between alternative equating designs and methods. In addition, equating involves many computations and the use of conversion tables, thus increasing the potential for mistakes.

*Quality control processes*

1. If there are unexplained equating problems (e.g., the common items that serve for equating do not have the same characteristics in the two test forms), confirm that the test was administered under the same standardized conditions (test format, time allowed, etc.)
2. Check that the specified equating procedures, the data collection design (e.g., using random groups) and the statistical method assumptions have been followed correctly.
3. Check that the correct conversion equation has been used.
4. Compare the scores obtained with prior expectations based on examinee background, exam date, and repeater data. If there is a discrepancy—investigate the reason for it.
5. If there are cut scores—check the pass and fail rates. Compare them to your prior expectations.

8. **Computing standardized scores**—Using parameters or conversion tables to compute the scale, standardized, or percentile score to be reported

*Explanation*
Final scores are reported on specific scales (SAT—from 400 to 1600, ACT—from 1 to 36 etc.) Usually there is a need to convert raw scores (number correct or number correct adjusted for guessing) or theta scores (IRT-based tests) to the test-specific scale. The conversion is done by means of a numeric table or a function (e.g., for linear transformation). Sometimes a "doglegging procedure"[3] is applied for defining a uniform minimum and maximum in each reported score.

*Mistakes or indications of the need for quality control—[Examples]*

1. Wrong conversion parameter used in the transition from raw/theta score to reported scores (e.g., a multiplier and an addend for linear conversion).
   - 1.1 A parameter that belongs to another test is used.
   - 1.2 There is a fault in the computer program that calculates the scores. For example, temporary default parameters are used instead of the operational parameters.
2. The wrong conversion table is used in the transition from raw/theta score to standardized score.
   - 2.1 One row in the table has a mistake in it.
   - 2.2 An entire column in the table is accidentally reversed.
   - 2.3 A conversion table that belongs to another test is used.
3. Rounding problems
   - 3.1 There are several computer programs that compute scores; one rounds 1.5 up to 2.0, the other rounds 1.5 down to 1.0.
   - 3.2 A temporary calculation (that should not be rounded) is rounded.

*Quality control processes*

1. Check low standardized scores to make sure that they are based on low raw scores.
2. Compare the tables/parameters of the new form to other test form tables/parameters; when test forms are parallel, the tables/parameters should be similar).
3. Disable table editing (tables should be "read only" files).
4. Calculate some scores manually and compare results with the computer-generated results.
5. Check the statistical relation between raw scores and standardized scores; examine a scatter plot that presents this relation.

9. **Test security checks**—to detect fraud

*Explanation*
Before scores are reported, ensure (to the best of your ability) that each score represents the achievement of each examinee. Reporting a score that was obtained by dishonest means is a serious problem. Cheating cannot be fully prevented since the temptation to cheat, especially in high-stakes testing, is great. Combating cheating also has legal aspects, which

should be taken into consideration (for a comprehensive review of this topic, see Cizek, 1999). In high-stakes national testing, fraud might be perpetrated on the class, school, or district level, as teachers and principals may be penalized for low scores and rewarded for high scores.

Sometimes, these checks reveal problems in data collection or storage rather than fraud.

*Examples of cheating*

1. Impersonation—An individual other than the "real" examinee takes the test.
2. Copying—An examinee copies answers from another examinee.
3. The test items and answers are known in advance.
4. The examinee communicates with an external person generally by electronic means.
5. Prohibited materials are brought into the test.
6. Help is received from test proctors during the test.

*Quality control processes—Actions to detect and prevent fraud*

Do your best to penalize cheaters. Have a documented legal routine for dealing with them. Inform examinees, in advance, that you take steps to combat fraud. Catching and punishing examinees can decrease general motivation to perform fraudulent acts.

*Detecting/preventing copying*

1. Do not seat two examinees who may be acquainted near one another.
2. Check aberrant or unexpected response patterns, e.g., if difficult items are answered correctly and easy items incorrectly.
3. Apply copying indices, based on the similarity of the answer sheet to those of other examinees in the same class (additional info: examinee location in class, previous scores). The most popular copying index is the K-index (see, for example, Sotaridona & Meijer, 2002)

*Detecting/preventing impersonation*

1. Validate picture ID as examinees are admitted into the classroom.
2. Obtain a handwriting sample from each examinee.
3. Analyze divergent repeater scores. Extreme differences may be the result of impersonation.
4. Monitor the test proctors.

Many more examples are provided by Cizek (1999, Table 9.1, p. 165).

10. **Reporting test scores**—Delivering scores to examinees (examinee reports) and to the client institutions (institution reports)

*Explanation*

Scores are reported *both on a printed form and electronically. Use of the Internet for score reporting is increasing.* Reporting must be done in such a way that people understand the meaning of their scores.

*Mistakes or indications of the need for quality control—[Examples]*

1. Examinees do not understand the meaning of their scores.
2. Individual score reports are forged.
3. Errors occur in the testing institution's reports.
4. Scores that should have been sent to one institution are sent to another institution.
5. A hostile individual obtains access to examinees' scores (hard copy or computer file).

*Quality control processes*

1. Use focus groups of examinees to construct a meaningful explanation of the score report.
2. Inform institutions that only the institution report and not the examinee report should be used.
3. Ensure that the individual score report is difficult to forge.
4. Never use a "computer editor" while preparing the institution report.
5. Encrypt electronic score reporting files.

11. **Documentation of the scoring process**

*Explanation*

The entire scoring process, including main statistics and trends over time, must be documented routinely throughout the process and completed not long after test scores are released. According to Rhodes and Madaus (2003), companies that regularly audit results are more likely to detect and correct errors.

*Mistakes or indications of the need for quality control*

1. Not documenting the process or documenting only certain parts of it.
2. Performing documentation a long time after test has been scored and results delivered.

*Quality control processes*

The current "documentation culture" has an effect on the SER process and will be significant in making future SER processes more reliable and accurate.

1. Routine documentation of the entire scoring process, including main statistics and trends, is critical.
2. Highlight extreme statistics that must be checked before scores are reported (e.g., low correlation between raters).
3. Do not deliver the new form before the old one has been documented.
4. Release some of the technical data and enable the public to contact you directly.

## Lessons From Other Industries

Different industries vary in their quality control processes and in the attitudes they adopt toward mistakes. In many cases, the majority of quality checks can be performed on the final product. (Is the car safe? Does the air conditioner work? Is the cake tasty?) However, when dealing with scores, almost all the quality checks must be carried out during the process and cannot be performed on the final product. Two examples of fields in which quality control and safety checks play an important role are the airline industry and software and hardware development. We have tried to learn from their experience. In software development, quality assurance is usually organized as a separate, independent group that acts according to standards and specified criteria. Fujii (1978) writes: "Independence provides a fresh viewpoint needed to accurately assess the software, and it also precludes bias in critiquing the software products" (p. 30). Independent quality control (done by professional/s who is/are not members of the operational staff) is vital in scoring, equating, and reporting.

## Testing Through the Internet

The Internet is increasingly employed for delivering, scoring and reporting test scores. It is an extremely useful medium that will certainly be put to more varied and extensive use in the near future. Most of the quality control procedures listed

in the module are also relevant to Internet-delivered testing. However, there are some special quality control procedures that such testing necessitates. The International Test Commission Guidelines on Computer-Based and Internet Delivered Testing (2005) is a useful and highly recommended source on the subject of quality control procedures. Two of the relevant topics in these guidelines are: (1) ascertaining the degree of control over the test conditions (very important when the purpose is to have all the scores on the same scale), and (2) giving consideration to control of test-taker authenticity and of cheating. This is usually easier in paper and pencil supervised testing.

**Concluding Comments**

In addition to the above recommendations, I would like to offer some general guidelines and recommendations. Each time a new test is introduced, a detailed simulation of the entire SER process should be carried out. (For an example, see Texas Education Agency et al. 2004). Then, standards for quality control checks should be formulated.

Furthermore, the human factor is vital for quality control: All individuals engaged in the SER processes should be professionals (e.g., equating experts) who are familiar with the written standards. Raters who evaluate open-ended items must participate in workshops and training sessions. They should be given rating instructions, sample papers, and subjected to assessment, before they begin evaluating examinee responses. Raters must meet predetermined standards and work independently.

Computers are also an integral part of the process. Hence, programs and interfaces must be reliable and user-friendly.

If something in the data appears strange, check it again and again. There is a reason for every mistake and anomaly. If you can identify the mistake, you can also identify the cause and prevent it from happening again. Needless to say, the mistake should be corrected immediately and the consequences for all examinees taken into account.

**Self-Test**

1. Choose three tests with which you are familiar. These might be MC (multiple choice), P&P (paper & pencil), open-ended P&P tests, and MC computerized tests. Formulate standards for the maintenance of accurate records of every examinee answer sheet. Try to write your answer in a table format.

2. Item analysis shows that a specific operational item is much more difficult (based on a delta index) than it was when it was administered as a pilot item. List three checks that are needed to verify this finding.

3. Study the following table, which presents statistics (mean and standard deviation in brackets) for a fictitious test over 3 consecutive years:

| | Year | | |
|---|---|---|---|
| Part | 2003 | 2004 | 2005 |
| Vocabulary | 320 (*97*) | 324 (*96*) | 337 (*91*) |
| Reading Comprehension | 307 (*101*) | 312 (*99*) | 313 (*100*) |

Answer the following questions:
   (a) Which finding in the table is not compatible with the other findings?
   (b) Try to provide at least three explanations for this finding.
   (c) Suggest possible checks to verify each explanation.

4. List the quality control processes for the reporting step.

5. True or false? Answer and justify.
   (a) Item analysis can detect a wrong item key.
   (b) Equating is based on statistical assumptions that should be confirmed prior to equating.

**Answers to the Self-Test**

1. Example of an answer:

| | Test | | |
|---|---|---|---|
| Topic | MC, P&P | Open Ended P&P | MC Computerized |
| Examinee ID | Make sure that the ID is correct—that the examinee showed up, that there is no other examinee with the same ID, and that the ID number does not conflict with other biographical data such as date of birth. | | |
| Accuracy | Maintain your scanners and OMRs routinely check the accuracy of a sample of your examinees manually. | | Do a simulation and make sure that every click on the keyboard is saved correctly. |
| Backups | Have a backup scanner or OMR; have a manual alternative for gathering examinee data. | | Have backup batteries; have UPS. |

2. Four possible checks:
   (a) Make sure that the item content has not changed.
   (b) Study the content of other items in the operational test for item dependency.
   (c) Compare the item delta to other difficulty indices (Pi, IRT b parameter) in the operational administration.
   (d) Review the pilot statistics. A computational mistake may have occurred.

3. Answers:
   (a) The Vocabulary score in 2005 is not compatible with the other scores (337, *91*).
   (b) There are three possible explanations: (1) a calculation error, (2) some of the Vocabulary items were leaked, (3) the 2005 population differs significantly from the populations in the two previous years.
   (c) The checks are: (1) verify the calculations using an independent check, (2) perform a DIF (differential item analysis) analysis— compare the performance of every item to the performance on the item in preceding administrations, (3) investigate the biographical data of the examinees, compare it to the 2003 and 2004 data; compare it to the data of previous years—2000, 2001, 2002 . . . if possible.

4. The quality control processes for reporting are:
    1. Use focus groups to construct a meaningful explanation of the score report.
    2. Inform institutions that only the institution report and not the examinee report should be used.
    3. Make the individual score report difficult to forge.
    4. Never use a "computer editor" while preparing the institution report.
    5. Encrypt electronic score reporting files.
5. True or false:
    (a) Item analysis can detect a wrong item key. This is generally true. The item analysis procedure can usually detect a wrong item key—an item that has a negative correlation with relevant criteria (and if it is an MC test, a "wrong" answer alternative has positive correlations with the same criteria). However, sometimes, in special cases, the item analysis looks fine even if the key is wrong.
    (b) Equating is based on statistical assumptions that should be confirmed after equating—False. Most of the statistical assumptions should be checked prior to equating. Some can be checked after equating.


## Acknowledgments

I would like to thank the National Institute for Testing & Evaluation (NITE). Most of my knowledge on this topic came from my work there. Special thanks go to my colleagues from the scoring and equating department, where the quality control procedures were developed and are routinely employed.


## Notes

[1]The following citation demonstrates the implications of the abovementioned pressure. It is attributed to Joanne M. Lenke, the former president of Harcourt, in referring to a huge error of judgment that affected some 257,000 examinees in 1999. "[The error] might have been caught if the company had more than two days to analyze data from 4.3 million test forms . . . before Wednesday's deadline for passing results in the Internet (Colvin & Groves, 1999; Rhoades & Madaus, 2003).

[2]One example of an essential quality control procedure in the construction process is determining that no two items in a test form are identical.

[3]The doglegging procedure is used when the equating procedure produces raw-to-scale conversions such that the scale score does not fit exactly to the unified scale range. This procedure adjusts the raw-to-scale conversion toward and at the two ends to ensure that the minimum raw score will be equal to the minimum scale score, and that the maximum raw score will be equal to the maximum scale score.

## References


Colvin, R. L., & Groves, M. (1999) State's students gain moderately in reading education. *Los Angeles Times,* Thursday, July 1, 1999, http://www.onenation.org/9907/070199a.html.

Fujii, M. S. (1978). A comparison of software assurance methods. *ACM SIGMETRICS Performance Evaluation Review*, *7*, 27–32.

International Test Commission (ITC) (2005). International Guidelines on Computer-Based and Internet Delivered Testing. http://www.intestcom.org/itc_projects.htm.

Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the K-index for detecting answer copying. *Journal of Educational Measurement*, *39*, 115–132.

Texas Education Agency, Pearson Educational Measurement, Harcourt Educational Measurement, & Beta, Inc. (2004). *Quality control procedures. Texas Student Assessment Program*. Chapter 9: Technical Digest (2003–2004). http://www.tea.state.tx.us/student.assessment/resources/techdig04/.


## Annotated References

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

This book describes ways to detect and prevent cheating. Various statistical methods for detecting copiers are explained.

(The W.) Edward Deming Institute web site.

This web site serves as a reference for the pioneering work and philosophy of W. Edward Deming (1900–1993) and others on quality control in the United States and in Japan. http://www.deming.org/index.html.

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

This book includes a long chapter dedicated to practical issues in equating (Chapter 8). This chapter discusses the importance of quality control in the equating process and lists some major quality control procedures.

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. http://www.asu.edu/educ/epsl/EPRU/documents/EPSL-0503-101-EPRU.pdf.

In the module context, understanding the motivation for corruption in high-stakes testing is an important step in quality control procedures that fight corruption (when it is relevant and possible).

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

The book documents human errors in scoring, equating, and reporting educational standardized tests that occurred over a period of three years, 1992–2002. Errors are divided into those detected by test developers and those detected by others.

Toch, T. (2006). *Margins of error: The testing industry in the No Child Left Behind era*. Washington, DC: Education Sector. http://www.educationsector.org/usr_doc/Margins_of_Error.pdf.

The effects of the NCLB Act of 2001 on standardized testing in the United States are discussed. Problems and the high cost of errors for the testing industry are described. The author recommends federal supervision to help prevent and detect errors.

Zapf, D., & Reason, J. (1994). Introduction: Human errors and error handling. In J. Reason & D. Zapf (Eds.), Errors, error detection, and error recovery. *Applied Psychology: An International Review, 43,* 427–432.

This is a short introduction to a special issue that deals with human errors. Some of the topics in the special issue are: errors in planning and in decision making, error prevention, detection and handling. All have relevance to quality control.