Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Spring 2017, Vol. 36, No. 1, pp. 58–68

Quality Control for Scoring Tests Administered in Continuous Mode: An NCME Instructional Module

Avi Allalouf*, *National Institute for Testing and Evaluation,* Tony Gutentag*, *The Hebrew University,* and Michal Baumer, *National Institute for Testing and Evaluation*

Quality control (QC) in testing is paramount. QC procedures for tests can be divided into two types. The first type, one that has been well researched, is QC for tests administered to large population groups on few administration dates using a small set of test forms (e.g., large-scale assessment). The second type is QC for tests, usually computerized, that are administered to small population groups on many administration dates using a wide array of test forms (CMT—continuous mode tests). Since the world of testing is headed in this direction, developing QC for CMT is crucial. In the current ITEMS module we discuss errors that might occur at the different stages of the CMT process, as well as the recommended QC procedure to reduce the incidence of each error. Illustration from a recent study is provided, and a computerized system that applies these procedures is presented. Instructions on how to develop one's own QC procedure are also included.

Keywords: computer-based testing, continuous mode tests, quality control, scoring

People will always err. We cannot change human nature but we can try to change the conditions under which we operate in order to make errors less likely (Reason, 1997). Testing agencies should not only be aware of errors, they should anticipate them. The cost of errors in testing can be high for stakeholders—examinees, testing agencies, and test users (e.g., academic institutions and companies that use assessments for purposes of hiring)—especially in the case of high-stakes exams. Moreover, testing agencies have an ethical responsibility to report accurate scores. Errors should be reduced and prevented through the use of adequate quality control (QC) procedures. Hence, QC in testing is paramount.

In recent years, concerns about QC in testing have increasingly been voiced by the public and other stakeholders. The academic literature reflects this trend, although it has followed the issue more slowly (AERA, NCME, & APA, 2014; Lee & von Davier, 2013). Most QC procedures to date have dealt with large-scale assessment, with little attention given to tests administered to small population groups, usually on many administration dates, using a relatively wide array of test forms (CMT—continuous mode tests). As tests become increasingly computerized, CMT becomes increasingly relevant. Therefore, the focus of the current ITEMS module is on QC for CMT. First, we discuss the similarities and discrepancies between traditional QC and quality assurance, in general and in testing in particular. We also review the scant academic research that has been done thus far on QC in testing for large-scale assessment and CMT. Second, we discuss errors in testing: the reasons for errors, the implications of errors, error management for the different stages of the testing process (testing process, as well), as well as the appropriate QC procedures with regard to each error. We then illustrate the use of QC for a CMT based on a recent study, and its computerization. Finally, we provide supplementary instructions on how to develop one's own QC procedures.

Quality Control

Quality control methods and procedures are used to ensure that a product or a process is fit to be used or implemented (Montgomery, 1996). QC was initially developed in disciplines where the cost of mistakes is very high, such as medicine and aviation, or in areas with a high likelihood of mistakes, such as software development. For the most part, QC procedures have been applied to products (e.g., manufactured goods and services), and can be applied to any area within a company or organization. Montgomery (1996) defined *quality* as inversely proportional to variability, and *quality improvement* as the reduction of variability in processes and products. Since variability may best be described in statistical terms, statistical methods play a central role in quality improvement efforts. Therefore, while quality control in manufacturing works on the assumption that the inspected products should be homogeneous, quality control in educational assessment works

*Avi Allalouf and Tony Gutentag contributed equally to this article. Avi Allalouf, Director of Scoring and Equating, National Institute for Testing and Evaluation, Jerusalem, Israel; avi@nite.org.il. Tony Gutentag, PhD student, Department of Psychology, The Hebrew University, Jerusalem, Israel; tonygute@gmail.com. Michal Baumer, Computerized Test Unit, National Institute for Testing and Evaluation, Jerusalem, Israel; michalb@nite.org.il.

on the assumption that individuals—as represented by their scores—are heterogeneous.

Deming (1982) saw quality control as something that results from prevention of errors by means of process improvement, not post hoc inspection. Post hoc inspection typically occurs too late in the process, is too expensive, and is often ineffective. Therefore, nowadays we acknowledge the importance of *quality assurance*. Quality is something that has to be "assured" throughout the entire work process and not only inspected at the end of the line (Reason, 1997). This module deals with quality control in the broadest sense, that is, in terms of both quality assurance for the process and post hoc inspection.

Quality Control in Testing

In any production process there is considerable potential for error, and the assessment process is no exception. In the fields of psychological and educational measurement, in order to produce a score a multiple-stage process takes place, in which each stage is heavily dependent on previous stages, and in which errors can occur at any point (Rhoades & Madaus, 2003). Therefore, QC must be implemented separately for each stage of the assessment process, beginning with test development, right through test administration and test analysis, and ending with test scoring, reporting and validation (see Quality Control for the Different Stages of the Testing Process and Table 1 for further details).

When assessing the quality of a test, the determining factor is whether it meets its objective, that is, whether the quality of the decision made on the basis of the test is satisfactory (Roorda, 2008). The International Test Commission (ITC) (2014) defined *QC in testing* 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, thereby minimizing error and enhancing measurement reliability. Maintaining and improving quality is achieved with the help of standards, codes, and guidelines (Muniz & Bartram, 2007; Roorda, 2008). These are developed by associations such as the American Educational Research Association (AERA), the American Psychological Association (APA), the National Council on Measurement in Education (NCME)—who collaborated on the *Standards for Educational and Psychological Testing* (2014)—and also by the European Federation of Psychologists' Associations (EFPA; see Bartram, 2001), and the International Test Commission (ITC, 2014). The standards highlight the importance of quality control in scoring: "Those responsible for test scoring should establish and document quality control processes and criteria. . . . The quality of scoring should be monitored and documented" (AERA, NCME, & APA, 2014, Standard 6.9, p. 118).

Quality Control for Continuous Mode Tests

Quality control procedures for test scores can be divided into two types. The first type is QC on tests that are administered to large population groups, usually on few administration dates using a relatively small set of test forms (i.e., large-scale assessment); for example, the College Board's paper-based SAT test, which is used for college placement, and is offered seven times annually. In 2015, a total of 1,698,521 examinees took the exam (College Board, 2015), which means an average of approximately 240,000 examinees per administration. Currently, QC in testing is performed mostly on this type of test, for which there are widely accepted methods (Allalouf, 2007; Kolen & Brennan, 2004). The second type, which is developing as exams become increasingly computerized, is QC on continuous mode tests (CMT), where tests are administered to small population groups, usually on many administration dates and locations, using a relatively large set of test forms. For example, in 2014–2015, the Educational Testing Service (ETS) computer-based GRE test used for admissions to graduate studies in various universities was administered to 576,220 examinees at more than 1,000 locations in over 160 countries (ETS, 2015). Thus, there was an average of approximately 580 examinees per administration.

Several attempts have been made to develop QC for testing, both large-scale assessment as well as CMT. With regard to large-scale assessment, control charts have been quite prevalent. A control chart makes it possible to track a control variable along the time axis. Deviations that fall outside a certain range, a systematic pattern that appears on the chart, or points that fall far short of expectations are referred to as being "out of control". The first type of control chart is the p-chart, which was used by Savic (2006) to evaluate grading processes in higher education. Another type is the cumulative sum (CUSUM) control chart (see Omar, 2010 for a review), used mostly for person-fit index in the computer adaptive testing environment. A third and widely used control chart is the Shewhart control chart. Omar (2010) used Shewhart control charts to ensure quality of the measurement process for rating performance items in operational assessment. Schafer, Coverdale, Luxenberg, and Jin (2011) also described using Shewhart control charts to monitor quality in a large-scale assessment program for elementary and middle-school children in Maryland. Lee and von Davier (2013) reviewed and examined the effectiveness of CUSUM and Shewhart control charts, as well as reviewing additional statistical methods, such as change-point models and hidden Markov models. Lu and Yen (2014) used longitudinal regression to identify scoring errors in a large-scale K-12 testing program over 5 years. In a recent paper, Sinharay (2016) demonstrates how tests for a change point can detect a change in a mean score, person misfit, and item preknowledge.

It is difficult and sometimes impossible to use QC methods of large-scale assessment for CMT, due to the small number of examinees in each administration and the large number of test forms used on each administration, as well as the short time span between test administrations and between a test administration and the reporting of scores. Consequently, it is necessary to carefully adapt large-scale assessment QC for use in CMT, and to develop specific QC for CMT. With regard to QC for CMT, research to date is limited. Lee and Haberman (2013) used a harmonic regression in order to monitor scores of very frequent and unevenly spaced administrations of an educational test that had a relatively short history of stable operation. Gutentag et al. (2013) used Shewhart control charts to monitor scores of a continuous mode test (for further details, see An Example of Quality Control on Continuous Mode Tests).

Appendix Table A1 summarizes and elaborates on the QC methods mentioned above (see Lee & von Davier, 2013 for an extensive review of most of these methods in the context of large-scale assessment). Note that while most of these techniques were used in large-scale assessment, they can also be used in CMT, with proper adaptations. Further research

Table 1. Errors and QC in the Different Stages of the Testing Process—Test Administration and Test Analysis

Assessment Stage	Substage	Monitoring	QC Procedures and Desired Outcomes for Tests Administered in Continuous Mode	Check Every:
Test administration	Test allocation	Allocation of test	**Desired**: Totally random or randomized with weights, as planned. **QC**: Checking that allocation of test forms to examinees is as planned. Checking to see if a repeat examinee received a different test form than the one received in their previous administration(s).	Month
		Order of test sections	**Desired**: Order (fixed, random, model-based, etc.) as planned. **QC**: Checking that the order of test sections is as planned.	Month
		Order of test items	**Desired**: Order (fixed, random, model-based, etc.) as planned. **QC**: Checking that the order of test items is as planned.	Month
	Testing conditions	Suitable test conditions	**Desired**: Test conditions are suitable. **QC**: Documenting test conditions and highlighting extreme cases. There is no relationship between the test conditions and the scores.	Month
		Problems during test administration	**Desired**: No problems; problems that arise are solved. **QC**: Having an active professional helpdesk for every test administration	Online
	Database	Accuracy of data saved	**Desired**: Obtaining the examinees' answers, and storing examinee data in a database. **QC**: Data cleansing, including verifying that the answer key is correct.	Month
Test analysis and scoring	Item properties	Item properties	**Desired**: Psychometric characteristics (discrimination indices, biserials, etc.) are as planned. **QC**: Performing item analysis, even for small samples, while keeping necessary reservations in mind.	Month
		Item stability	**Desired**: Item properties are stable. **QC**: Performing drift analysis and investigating divergent items.	3 months

(*Continued*)

Table 1. (*Continued*)

Assessment Stage	Substage	Monitoring	QC Procedures and Desired Outcomes for Tests Administered in Continuous Mode	Check Every:
	Calculating scaled scores	Transformation into scaled scores	**Desired**: Transforming of raw scores or IRT theta into scaled or standardized scores properly and in a stable manner. **QC**: Analyzing and plotting the relationship between raw scores and scaled scores.	3 months
	Group level	Test form equivalence	**Desired**: Test forms received by the examinee are equivalent, and observed differences can be explained (e.g., demographic variables). **QC**: Compare the scores of different test forms using analysis of variance (ANOVA). Compare scores obtained with those anticipated on the basis of the test taker's background. Check that score distribution and score statistics are consistent with those observed in the past. If there are discrepancies, recheck the scores.	Month
		Stability of scores over time	**Desired**: No significant deviations. If deviations exist, interpretation is needed. **QC**: Monitoring scores over time (days, months, weeks) to track deviations. Shewhart control charts are the recommended method.	Month
		Test hall mean scores	**Desired**: No extreme cases, or at least a logical explanation for each divergent mean. **QC**: Compare test hall means to one another, and to test hall achievements in previous years, highlighting and examining extreme cases.	Month
	Examinee level	Repeat examinees' performance	**Desired**: The score difference (gain or loss) is reasonable. **QC**: Identifying examinees whose repeat scores are significantly higher or lower than the scores on their previous exam(s). Summon for retesting when score is questionable.	Week
		Subscore difference	**Desired**: Testing whether the differences in subscores are reasonable for a certain examinee. **QC**: Identifying cases of extreme discrepancy between subscores in the same test.	Week

should directly examine the applicability of these methods to CMT as compared to large-scale assessment.

Errors and QC in the Different Stages of the Testing Process

There are two major types of errors in testing—random measurement errors (not addressed here) and nonrandom errors (Rhoades & Madaus, 2003). Nonrandom errors are usually cases of human error (Senders & Moray, 1991) and computerized error (Peterson, 1996), but can also be the result of other factors (e.g., nonstandardized test conditions). A variety of factors contribute to the incidence of errors. They occur because standards were not established, or standards were established but not documented, followed or updated properly; scheduling (e.g., urgency in reporting test scores), and/or budget constraints prevented proper execution of the scoring process; or human factors came into play (e.g., lack of coordination, communication, and accountability) (Allalouf, 2007).

Errors may impact badly on examinees, test users, and testing agencies. An error in scoring may prevent an examinee from registering in or getting accepted to an educational institution or program. As for test users, an error in scoring may lead to individuals being assigned to an inappropriate educational program, or to the granting of a professional license to a person who lacks the required qualifications, or to cases of misguided intervention. Errors may also impact on the testing agency, sometimes resulting in legal action being taken against it, and may lead to the loss of credibility of—and public confidence in—educational and psychological assessment (Allalouf, 2007; ITC, 2014). Apart from the problem of the adverse consequences created by errors in testing for examinees, test users and testing agencies, there is an ethical dimension to the issue. Testing agencies have a responsibility to develop and implement reliable and valid tests with a minimum of errors (AERA, NCME, & APA, 2014). Therefore, QC of the assessment process is highly important.

Error management (EM) is aimed at dealing with errors and reducing the incidence of errors. It is indistinguishable from quality management. Most attempts at error management are sporadic and reactive; adequate EM, however, should be deliberate and planned. EM has two components: *error containment* and *error reduction*. While error containment is designed to limit the adverse consequences of errors that have occurred, error reduction entails measures designed to prevent or limit the occurrence of errors (Reason, 1997). Error containment is treated using the *error handling process* (Zapf & Reason, 1994). In cases where an error has occurred, it has to be diagnosed (e.g., detected and explained). From an organizational point of view, it is important that there be a high rate of error detection. However, identifying an error is merely the beginning of the search for an explanation; understanding the context that gave rise to the error can hopefully limit its recurrence, thereby contributing to error reduction efforts (Reason, 1997). After the diagnosis has been completed, error recovery should take place (Zapf & Reason, 1994).

Quality Control for the Different Stages of the Testing Process

Educational measurement through testing involves five main elements: test development, test administration, test analysis, score reporting, and validation. This module focuses on three of these elements: test administration, test analysis and score reporting. The possible errors that might occur during test administration and test analysis are presented in Table 1. Suitable QC procedures, specifically relevant for CMT, are suggested for addressing each error. The score reporting stage will also be discussed briefly.

In Table 1, the two stages of test administration and test analysis are each divided into substages that are organized chronologically. For each of the substages, a "desired" criterion is defined that characterizes an "in-control" testing process, followed by a QC procedure aimed at meeting this criterion by preventing and/or containing errors.

Test administration. Making sure that the test administration was carried out properly is extremely important when assuring the quality of an exam; otherwise, it means that the score or ability of the examinee was not measured reliably and validly. This stage is divided into three substages. The first substage is test form allocation. What should be monitored using QC procedures in this substage is whether the allocation of test forms for examinees, the order of test sections in a given test form, and the order of test items in a given section—are all as planned. If this is not the case, the success or failure of an examinee could be attributed to a test form or to a section or item order effect instead of to his/her ability. QC at this stage is especially important in CMT, which is mostly computerized, meaning that test allocation (forms, sections, and items) is more complex.

The second substage focuses on test conditions. What should be monitored is whether test conditions are suitable, and whether problems that arise during test administration are solved. If there are problems in either case, the success or failure of an examinee could be attributed to test conditions instead of to his/her ability. QC at this stage is especially important in CMT, because test conditions tend to vary more when administration is computerized and can potentially take place at any location. The third substage focuses on the database. What should be monitored and ensured is that all of the data is collected and properly stored and sorted, because the quality of output is determined largely by the quality of the input. Otherwise, there is the danger of GIGO: "garbage in, garbage out". QC at this stage is especially important in CMT because data are treated automatically without further inspection.

Test analysis and scoring. Quality control at the test analysis and scoring assessment stage, which is the heart of the scoring process, is highly important. This stage is divided into four substages. The first substage deals with item properties. What should be monitored using QC procedures in this substage is whether items' psychometric characteristics are stable and as planned. If this is not the case, it means that some items will be unreliable and invalid, and examinees will receive items that are not optimal for assessing their ability. The second substage is calculating scaled scores. What should be monitored and ensured is that raw scores or ability levels be transformed properly into scaled scores and in a stable manner (e.g., using conversion tables and monitoring the use of these tables); otherwise the score will be wrong. The third substage is examining the scores at the group level and ensuring that test forms are equivalent, that test scores are stable over time, and that mean test hall scores are reasonable; a problem in these areas might be a sign of fraud. The

fourth substage is checking the scores at the examinee level, which helps detect cases of cheating; this is done by using repeat examinees' data and looking for extreme differences as compared to previous administrations or among subscores in different test sections. Differences in examinee level may also indicate that there is a problem in the collection of data. QC for CMT in this stage should focus on comparing the administration mean (e.g., test score, psychometric characteristics of items), even for very small samples, to long-term trends and to a known benchmark (e.g., other forms, test hall mean, previous administration of the examinee and/or the examinee's performance in a different section, mean administration score on the same month of the previous year). The small sample size in CMT may induce false alarms, but "better safe than sorry."

Score reporting. A central QC issue relates to the score reporting stage. Nowadays, in testing in general and in CMT in particular, especially when the test is computerized, there is an expectation or even demand that scores be reported as quickly as possible (within a few days or even immediately). On the one hand, the swift reporting of a score reduces stress for the examinees, augments their feeling of satisfaction with the testing process, and of course provides them and other test users with the desired information (i.e., a score). On the other hand, the short time frame entailed in swift score reporting may impinge on the thoroughness with which QC procedures on the scores are conducted.

One solution to this quandary is to report a tentative score to examinees online, with further in-depth QC procedures done post hoc that may change the reported score to a different, more accurate, final score. Another solution is to report scores with minimal delay and conduct all necessary QC procedures in the meantime (revising the score if necessary). One way or the other, we not only advise against leaving out the essential stage of QC on scoring; we recommend doing it thoroughly, despite the serious pressure from examinees and other test users. An efficient and structured QC procedure that is prepared in advance can be helpful in reducing the time gap between test administration and the reporting of scores. An automated QC system can help in reducing this gap even further.

An Example of Quality Control on Continuous Mode Tests

The following is presented by way of illustration of QC procedures for continuous mode tests as outlined above, using real test data. The research was conducted on an online computerized CMT test—MEIMAD—which is offered by the National Institute for Testing and Evaluation (NITE) for preacademic preparatory program placement. The test is offered approximately 120 times annually, and about 6,000 examinees take the exam annually (an average of 50 examinees in each administration; ranges from approximately 5 to 150 examinees). The research was based on data pertaining to 24,548 examinees who took the test over a period of 4 years, from 2008 to 2012 (Gutentag et al., 2013).

The MEIMAD is a multiple-choice test used in preacademic preparatory programs in Israel (Baumer, Gafni, Turvall, & Schweid, 2010). Over the past few years, approximately 6,000 examinees have taken the test annually. The test has eight sections, six of which are operational (two for each test domain: mathematics, verbal, and English) and two of which are pilot sections. Administration of the test has been computerized since 2008, and takes place at various educational institutions throughout the country on many administration dates and for small population groups. The relevant QC procedures presented in Table 1 were applied to the data.

First, the test administration stage was examined. On each administration date, one test form is randomly selected, and in each of these test forms the sections are presented in a random order. A chi-square analysis of a weighted sampling of test forms found that the test form weight was as planned ($\chi^2\,[7, N = 4{,}452] = 8.94, p = .261$). Moreover, within each test form, the order of sections was examined using a chi-square analysis and was also found to be random ($\chi^2\,[49, N = 10{,}720] = 48.68, p = .486$). Second, the percentages of correct answers in a given form in 2009 were compared

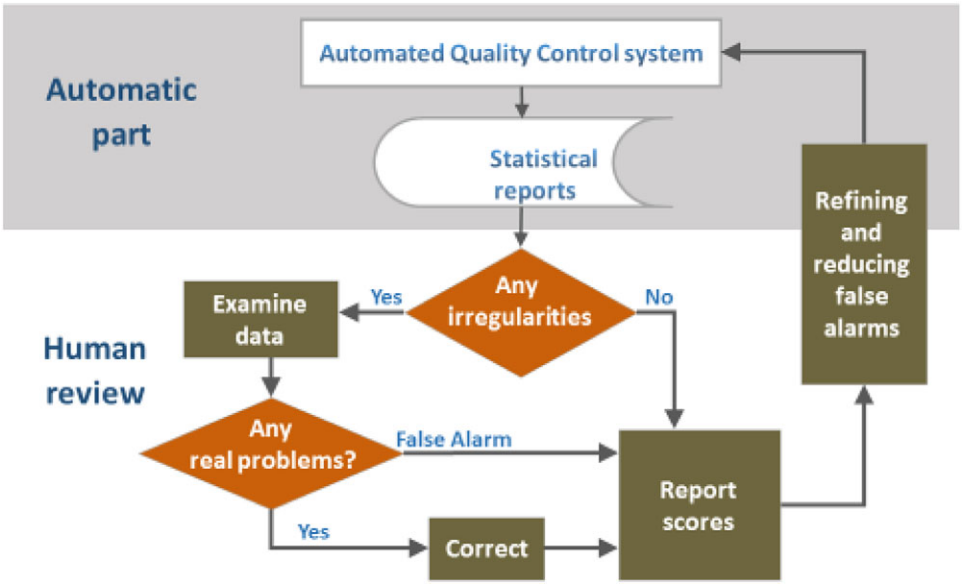


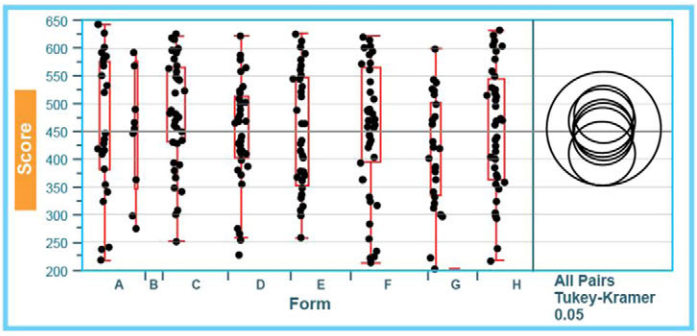
FIGURE 1. Automated quality control system—a flow chart. [Color figure can be viewed at wileyonlinelibrary.com]

Weekly Summary

workweek	from_date	to_date	N
2015_W37	20150913	20150920	289
workweek	**language**		
2015_W37	Hebrew		272
	Arabic		17
date	**campus**		
16Sep2015	AAC		79
	MVC		17
	OBC		58
17Sep2015	HU		36
	LEC		31
	THC		14
	EMC		54
gender			
Missing			21
M			124
F			144
language	**extra time**	**in domain**	
Hebrew	1	All	172
	1.25	All	75
		VE	25
Arabic	1	All	17

1

Scores by Test forms, Language=Hebrew



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Form	7	93429.3	13347.0	1.1598	0.3263
Error	264	3038186.3	11508.3		
C. Total	271	3131615.5			

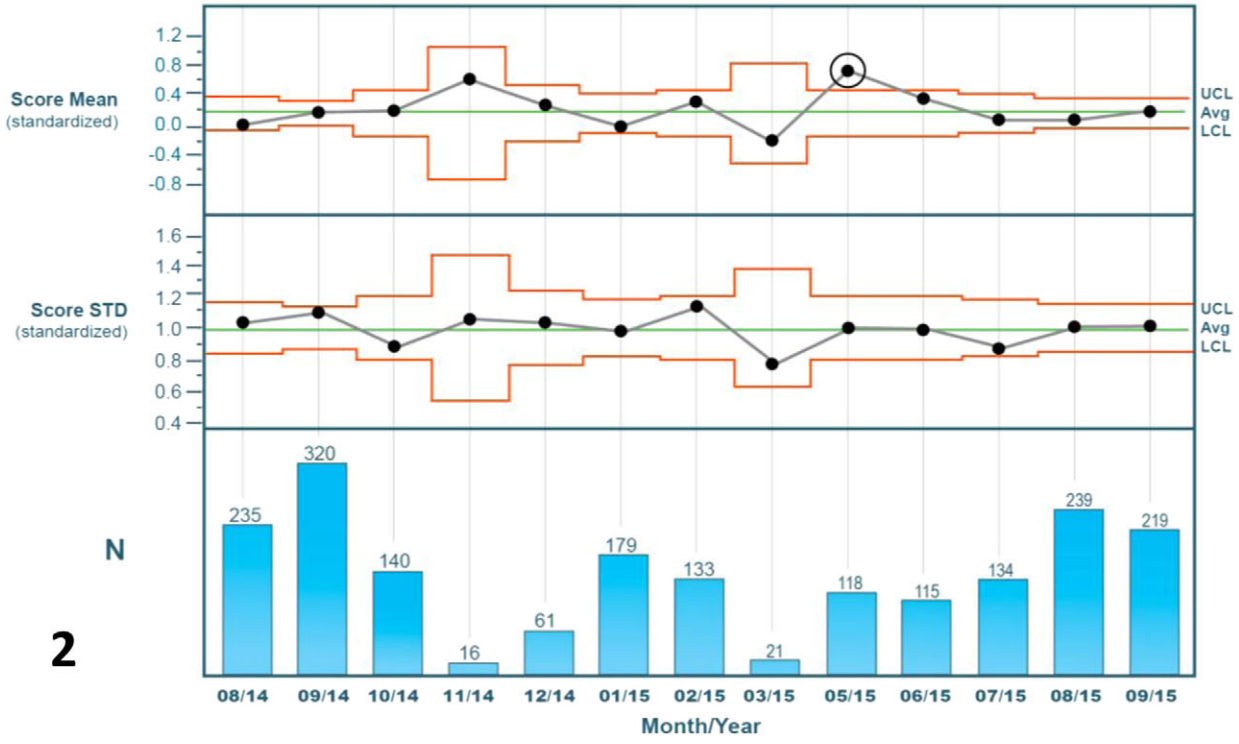
Means Comparisons

Comparisons for all pairs using Tukey-Kramer HSD

Connecting Letters Report

Form		Mean
C	A	482.22500
A	A	465.18519
F	A	456.59574
B	A	455.20000
H	A	454.22727
D	A	446.92105
E	A	443.69444
G	A	412.16667

3



2

FIGURE 2. A screenshot of the computerized QC system. [Color figure can be viewed at wileyonlinelibrary.com]

with those in 2011; it was found that item difficulty did not change over the years (e.g., in the verbal domain $r = .99$, $p < .001$; and the mean score in 2009 and 2011 was 60 and 59, respectively). Third, analysis of variance (ANOVA) was used to examine test form equivalence. A significant difference was found among test forms, but effect size was miniscule (F [8, 24,539] = 12.52, $p < .001$, $\eta^2 = .04$). In a post hoc examination, one particular test form was found to be significantly lower ($M = 385$) from the other test forms ($M =$ 400 to 415) in the total score, supporting a previous decision to stop using it. Fourth, the scores of repeat examinees were highly correlated (Mean $r_{\text{test–retest}} = .78$). Next, we examined the data of the continuous mode test in a more continuous manner. In order to examine the stability of scores over time we used a Shewhart (1931, 1939) control chart. In the study, some deviations were found in a repeated pattern over the years. Part of the overall variance was explained by using a regression with relevant variables: the predictors of gender, time accommodation, and academic institution in which the exam took place ($R^2 = .18$ adjusted $R^2 = .18, F$ [48, 18,702] = 84.46, $p < .001$). When plotting a Shewhart control chart for the residuals, very few deviations remained over time (only in the months with extremely small sample sizes, April and November).

To conclude, application of QC methods for CMT can facilitate fast and accurate identification of errors at the different stages of the testing process. In the current investigation, no irregularities were found in the test administration stage: test form allocation was as planned, and the order of test sections in a test form was random as planned. With respect to the test analysis and scoring stages, (a) item properties, as examined by item difficulty, showed stability; (b) test form equivalence in the group level showed that one test form was significantly different than the other—it was removed from use, and had been revised; (c) scores in the group level showed stability over time; and (d) repeat examinee performance showed high correlation at the examinee level.

An Automated Quality Control System for CMT

Based on the study described above (Gutentag et al., 2013), an automated QC system was built, and is now being refined, that examines the administration data of the MEIMAD exam (Baumer, 2017). This system has two parts: an automatic part and a human review (see Figure 1).

Automatic part. The automatic part consists of data extraction and generating of reports. The system begins by extracting the newest data from NITE servers on a weekly basis. These data are added to a cumulative database saved separately for QC system analysis use. Then, using SAS programming, the QC system mimics the QC procedures mentioned above. As can be seen in the screenshot of the system presented in Figure 2, these procedures produce reports, statistics summarized in tables (marked as 1 in Figure 2), graphs (Shewhart control chart; 2), as well as tests of statistical significance (3).

Human review. Once the reports are produced, an expert—trained to recognize irregular data—reviews the summarized data prior to score reporting. If no errors are found in the scores, the reporting of the scores to examinees and test users proceeds as usual; but in cases where irregularities are recognized, an in-depth and careful exploration of the scores is executed (see Table 1 for details). If a real problem is evident, it must be corrected before scores can be released. If it turns out to be a false alarm, scores can be reported. In all cases, the process should be improved over the long run and the number of false alarms should be reduced.

As the fine-tuning of the QC procedure continues and the definition of what is regarded as irregular comes into focus, a dashboard summarizing the QC procedure's main findings and flagging irregularities is being developed. In future, we plan to expand the QC system so it will monitor additional tests using the same technology and the same QC procedures.

Build Your Own Quality Control Procedure

We believe that Table 1 offers a comprehensive list of errors. It may be, however, that having examined the list, you may be interested in creating your own tailor-made QC procedure based on the unique needs, processes—and past errors—of the assessment process in your organization. The following steps were designed to guide you through this process.

1. Write a comprehensive list of past errors in the test(s) you want to monitor, based on data from your own organization as well as other organizations similar to yours. You are advised to consult the relevant academic literature.
2. Map errors according to their chronology in the testing process, grouped by test elements (e.g., test administration and test analysis).
3. Match each error to the relevant QC procedures, that is, those that can help in identifying and/or preventing each error. Table 1 will be helpful in this, as will your own personal experience and/or the experience gained in your or others' organizations.
4. Run a pilot investigation on a test that was administered recently or is being administered now, according to the list of QC procedures you assembled during stages 1–3. Try to see if your proposed QC procedure is comprehensive and indeed suitable for identifying and/or preventing all possible errors you referred to in the testing process. Refine your proposed QC procedure according to your findings.
5. Develop an automated QC system that will operate before score reporting, and monitor the errors assumed to occur for the different test stages, according to the QC procedures you have outlined so far.
6. Plan timely meetings in your organization devoted to discussing QC procedures and their findings.
7. Document your QC procedures, findings, conclusions, reactions and the lessons learned after each run and each timely meeting.
8. Clearly specify whose responsibility it is to build, execute, and routinely monitor this QC process. It is recommended that a trained professional(s) be assigned the task of routinely monitoring QC, who will operate independently of those involved in routine test scoring, analysis and score reporting (ITC, 2014).

Concluding Comments

Quality control in testing is imperative. Quality should be ensured throughout the testing process and not only at the end

of the line. Errors in testing are to be expected but should be managed—reduced to a minimum whenever possible or contained when prevention fails. Despite the pressure from stakeholders to report scores as quickly as possible, testing agencies have an ethical responsibility to report accurate scores, and to ensure this by following the necessary quality control procedures. QC for CMT is a challenge, and not much research has addressed this important issue to date. In the current ITEMS module, we discussed errors that might occur in CMT at the different stages of the testing process, we presented the suitable QC procedure for each error, and we illustrated the latter using an example from a recent study. QC that relates to specific stages in the testing process can assist in identifying past errors, detecting current errors, and preventing future errors. A computerized QC system can facilitate the rapid and continuous monitoring needed for dealing with CMT, and thereby reduce the incidence of errors even further. We hope this module has been helpful in highlighting the importance of QC in testing in general and in CMT in particular, in indicating the kinds of QC procedures suitable for possible different errors, and in providing the groundwork for the eventual building of one's own QC procedure.

Self-Test

1. Table 3 represents Test A and Table 4 represents Test B. One of the tests is a CMT test and the other is a large-scale assessment. Complete the following tables using the data below, and identify which table represents which test type.

Table 3: Test A

		Number of Examinees	
Date	**Number of Test Forms**	**Total**	**Average per Form**
April 5	1	3,000	__
____7	__	9,000	4,500
____11	2	__	4,000

Table 4: Test B

		Number of Examinees	
Date	**Number of Test Forms**	**Total**	**Average per Form**
April 5	6	__	20
____7	7	210	__
____11	__	90	15

Data: April, June, July, April, 3,000, 120, 2, 6, 8,000, 30.

2. Name three differences between quality control for products and quality control for scores.
3. Specify the relationship between quality control and quality assurance.
4. Which of the following is not a type of control chart?
   a. Shewhart
   b. *p*-chart
   c. CUSUM
   d. HMM
5. What is common to the following three QC methods: Longitudinal regression, change-point model, and harmonic regression?
   a. They were employed only in large-scale assessment.
   b. They were employed only in CMT.
   c. They employ regression.
   d. They are types of control charts.
6. After a test was administered to a group of examinees, the testing agency discovered that a wrong answer key was used to compute the scores. The testing agency used the correct answer key to recompute the scores, and then reported the corrected scores to examinees and institutions. This is an example of:
   a. Error containment
   b. Error reduction
7. Go to Rhoades and Madaus (2003), and choose one of the errors described there. Explain how this error might have been prevented, using the appropriate QC procedure(s).
8. Because an automated quality control system can assist in examining errors in real time, only recent data are needed in order to decide whether a certain data point is "out of control" when plotting a Shewhart control chart for the mean. Yes/No

Answers to the Self-Test

1. The test type and the complete tables are presented below (completed dates and figures are shown in boldface).

Test A: CMT Test

		Number of Examinees	
Date	**Number of Test Forms**	**Total**	**Average per Form**
April 5	1	3,000	**3,000**
June 7	**2**	9,000	4,500
July 11	2	**8,000**	4,000

Test B: Large-Scale Assessment

		Number of Examinees	
Date	**Number of Test Forms**	**Total**	**Average per Form**
April 5	6	**120**	20
April 7	7	210	**30**
April 11	**6**	90	15

2. Three differences between quality control for products and quality control for scores, respectively, are the inspected object (product versus score [representing human ability]), the type of inspected object (homogeneous versus heterogeneous), and the aim of QC (reducing variability versus ensuring that variability is real and clear of noise).
3. *Quality control* is a systematic process designed to help ensure that high-quality standards are maintained at all stages of the production process (i.e., *quality assurance*) as well as inspected post hoc, at the end of the line.
4. (d) HMM is not a type of control chart.
5. (c) They employ regression. To date, these methods have been used in both large-scale assessment and in CMT, and are not a type of control chart.
6. (a) Error containment. The testing agency's steps were designed to limit the adverse consequences of an

error that had already occurred (use of a wrong answer key) and therefore, it is an *error containment* procedure. If the testing agency were to henceforth adapt a routine QC to check whether the right answer key is always used to compute the scores, then we would say that an *error reduction* procedure was being used.

7. Table 1 can assist you in finding the appropriate QC procedure(s) needed to prevent the chosen error.
8. No. In order to decide whether a certain data point is "out of control," it should be examined in comparison to long-term parameters, such as the mean and standard deviation of the score in the previous year(s), and/or compared to the parallel data points in previous years.

Appendix

Table A1. Quality Control Methods in Testing

Method	Description
Shewhart (1931, 1939) control charts	Shewhart control charts are sketched in two dimensions: time on the vertical axis, and observation on the horizontal axis. The graph consists of four lines: the baseline, which is generally the measurement of central tendency or predicted value; the variance for the control bounds, with the top and bottom lines being an equal distance above or below (respectively) the baseline (usually ± 3 *SD*); and a connecting line among the observed means of the scores for each administration time.
p-Chart	A *p*-chart is used to monitor the proportion of nonconforming units (e.g., that do not meet the desired criteria) in a sample. The chart displays the nonconforming ratio versus time in a similar manner to Shewhart control charts.
Cumulative sum (CUSUM) control charts	CUSUM charts are designed for the rapid identification of a permanent change. The measurement for a certain point in time is the accumulated deviation from a particular target value of the process in a similar manner to Shewhart control charts.
Hidden Markov models (HMM)	A general HMM involves two main components: the measurement model and the transition dynamics. Given a discrete state space, the measurement model captures the relationship between observation and states, whereas the transition dynamics characterizes the relationship between states over time.
Longitudinal regression	Longitudinal regression refers to the application of regression procedure, linear or nonlinear, to longitudinal data where the outcome of an individual is measured at several different points in time (e.g., achievements at different grade levels).
Change-point model	Change-point model is a regression with two aspects of the change-point approach to process monitoring: determining if there has been a change in the process and estimating the time of the change. This model is used to match models from different distributions to different periods of time in the data, when one is interested in finding the transition points between distributions.
Harmonic regression	In harmonic regression, the predicted variable(s) are trigonometric functions of time (and therefore, cyclical). This type of regression allows the creation of a model (pattern) for seasonal data.

Acknowledgment

We would like to thank the following people for their assistance with this project: Marina Fronton, Keren Roded, Nadav Podoler (all from NITE) and Sigal Levy from the Academic College of Tel Aviv Yaffo.

References

Allalouf, A. (2007). An NCME instructional module on quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice*, *26*(1), 36–46.

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

Bartram, D. (2001). Guidelines for test users: A review of national and international initiatives. *European Journal of Psychological Assessment*, *17*, 173–186.

Baumer, M. (2017). *An automatic quality control system* (Research Report). Jerusalem, Israel: National Institute for Testing and Evaluation.

Baumer, M., Gafni, N., Turvall, E., & Schweid, T. (2010). *Pre-academic preparatory program admissions tests (MEIMAD)—Summary report of internet administration in 2008–2009* (Internal Report No. 184). Jerusalem: National Institute for Testing and Evaluation. [in Hebrew]

College Board. (2015). *2015 college-bound seniors total group profile report*. Retrieved December 20, 2015, from: https://secure-media.collegeboard.org/digitalServices/pdf/sat/total-group-2015.pdf

Deming, W. E. (1982). *Out of the crisis*. Cambridge, MA: MIT Press.

Educational Testing Service (ETS). (2015). *A snapshot of registration, test centers and dates*. Retrieved on April 17, 2016 from http://www.ets.org/gre/revised_general/register

Gutentag, T., Allalouf, A., Baumer, M., Levy, S., Fronton, M., & Roded, K. (2013, October). *Quality control for scoring continuously administered tests*. Paper presented at the 39th Annual International Association for Educational Assessment Conference, Tel Aviv, Israel.

International Test Commission (ITC). (2014). ITC guidelines on quality control in scoring, test analysis, and reporting of test scores. *International Journal of Testing*, *3*, 195–217.

Kolen, M. J., & Brennan, R. L. (2004). Practical issues in equating. In M. J. Kolen & R. L. Brennan, *Test equating: Linking and scaling* (pp. 268–328). New York, NY: Springer.

Lee, Y., & Haberman, S. J. (2013). Harmonic regression and scale stability. *Psychometrika*, *78*, 815–829.

Lee, Y., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika*, *78*, 557–575.

Lu, Y., & Yen, W. M. (2014). Use of longitudinal regression in quality control. *ETS Research Report*, *2014*(2), 1–20.
Montgomery, D. C. (1996). *Introduction to statistical quality control* (3rd ed.). New York: John Wiley.
Muniz, J., & Bartram, D. (2007). Improving international tests and testing. *European Psychologist*, *12*(3), 206–219.
Omar, M. H. (2010). Statistical process control charts for measuring and monitoring temporal consistency of ratings. *Journal of Educational Measurement*, *47*, 18–35.
Peterson, I. (1996). *Fatal defect: Chasing killer computer bugs*. New York, NY: Vintage books.
Reason, J. (1997). *Managing the risks of organizational accidents*. Aldershot, UK: Ashgate.
Rhoades, K., & Madaus, G. (2003). *Errors in standardized tests: A systemic problem* (NBETPP Monograph). Boston, MA: Boston College, Lynch School of Education. Retrieved February 6, 2017, from http://www.bc.edu/research/nbetpp/statements/M1N4.pdf
Roorda, M. (2008). Quality systems for testing. In C. L. Wild & R. Ramaswamy (Eds.), *Improving testing: Applying process tools and techniques to assure quality* (pp. 145–176). New York, NY: Lawrence Erlbaum.
Savic, M. (2006). *p*-Charts in the quality control of the grading process in the high education. *Panoeconomicus*, *3*, 335–347.
Schafer, W. D., Coverdale, B. J., Luxenberg, H., & Jin, Y. (2011). Quality control charts in large-scale assessment programs. *Practical Assessment, Research and Evaluation*, *16*(15), 1–7.
Senders, J. W., & Moray, N. P. (1991). *Human errors: Cause, prediction and reduction*. Hillsdale, NJ: Lawrence Erlbaum.
Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. New York, NY: Van Nostrand.
Shewhart, W. A. (1939). *Statistical method from the viewpoint of quality control*. Washington, DC: U.S. Department of Agriculture.
Sinharay, S. (2016). Some remarks on applications of tests for detecting a change point to psychometric problems. *Psychometrika*, 1–13.
Zapf, D., & Reason, J. T. (1994). Introduction: Human errors and error handling. *Applied Psychology: An International Review*, *43*, 427–432.