GMAC®

Graduate
Management
Admission
Council®
Creating Access to Graduate Business Education℠

Evaluating Scale Stability of a Computer Adaptive Testing System

Fanmin Guo and Lin Wang

GMAC® Research Reports • RR-05-12 • November 30, 2005

Abstract

Scale stability is an important quality for any large-scale computer adaptive test (CAT) program and should be maintained through research on scale drift evaluations in the CAT operations. However, there is scarcely any literature on evaluating scale drift with CAT using both observed and simulated data. A method for evaluating scale drift is outlined and illustrated in this paper. In this study, a special online data collection method for the GMAT® Quantitative measure was designed and implemented. A modified root mean squared difference statistic was used to measure the difference in item parameters. Then an empirical baseline was established using simulations for evaluating the difference. The result showed that scale drift was not detected in the GMAT® Quantitative measure and the observed differences between the two sets of item parameters calibrated at two time points were random variations.

Introduction

For a testing program that has multiple administrations with new test forms per year over many years, it is critical to maintain a stable reporting score scale so that scores are comparable across administrations and test forms. Special studies are conducted by psychometricians from time to time to monitor the stability of a test's reporting score scale. In a linearly administered test such as a paper-pencil test (PPT), every examinee sees the same test items in the same test form during an administration. The test form used in a particular administration is equated to a reference form. The focus of a scale stability study is to identify scale drift as a result of equating a new test form to "one or more of the existing forms for which conversions to the reference scale (i.e., the reporting scale) are already available" (Angoff, 1984, p. viii).

In a computer adaptive test (CAT), however, each examinee can potentially see a different collection of operational test items that are selected from a large pool of calibrated test items, and there is not a unique test form for all examinees. The focus of monitoring scale stability is to see if there is any scale drift resulting from errors in the process of item calibration and parameter scaling of new items over time. This is because CAT operates within the framework of the item response theory (IRT) and employs some online item calibration method to calibrate and scale new items (Glas, 1998). In the IRT framework, all examinees' ability estimates and item parameters share the same scale, called the θ scale by convention. In an online item calibration design, new items can be linearly administered together with operational items that are administered adaptively and can be calibrated and scaled such that the new items are put on the existing scale of the operational items.

Errors in the calibration and scaling process may lead to certain degrees of scale drift over time, and this, in turn, may impact the test scores to the extent that, in the worst scenario, scores may not even be comparable. In a CAT program, new items are developed, calibrated, and added to an item bank from time to time, and old items are retired due to over-exposure or other reasons. Although efforts have been made to maintain a stable scale over time, where new items are constantly calibrated and put on the same scale with the operational items that have already been calibrated, this is not a guarantee for scale stability. In fact, accumulated errors in the calibration and scaling process can lead to a scale change. When this happens, the original interpretation of scores may no longer be valid. Therefore, it is both important and necessary to monitor scale stability over time in a CAT program.

Although extensive research has been conducted on various equating methods for linear tests like PPT (Kolen & Brennan, 1995), little can be found in CAT literature on scale stability issues. Prior to a study by Guo and Wang (2003), there was only one directly related reference by Stocking (1988). In that study, Stocking conducted a sequence of six rounds of simulations for online calibration in a CAT environment. All the data were simulated and Stocking did find evidence of scale drift from her study.

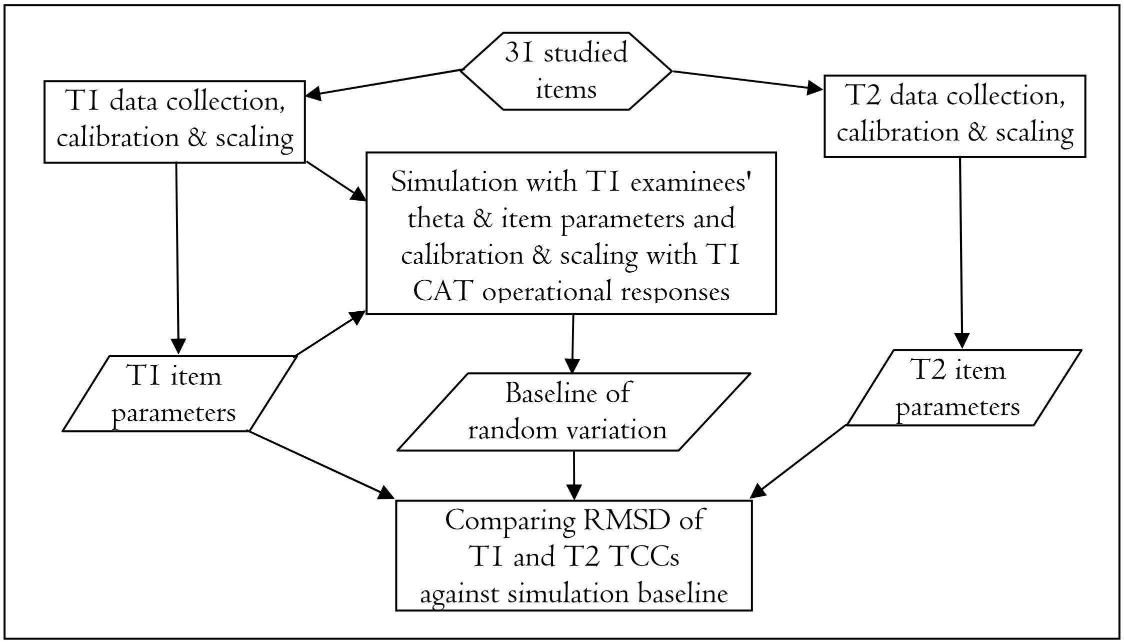
As CAT becomes mature and is used more often, it is essential to call attention to scale stability in CAT programs and design appropriate research to monitor and control scale stability in CAT programs by using both simulations and, more importantly, observed data from live CAT administrations. This was the very intent of designing and conducting this study.

The purpose of this study was to demonstrate design and analysis methods to examine the scale stability of a large-scale operational CAT program by investigating the calibration and scaling of test items that were pretested online. GMAT® quantitative CAT measure data were used in this study. Specifically, using both real and simulated data, the study was to accomplish three objectives: (1) Develop a special online data collection method to study the scale stability, (2) Demonstrate a method of using simulations to establish an empirical baseline for evaluating differences, and (3) Apply this method to the GMAT® quantitative CAT measure for an evaluation of its scale stability.

Methodology

Figure 1 is a flowchart of the design for this study. The design started with the selection of 31 studied items (the items of interest in this study). Two waves of data were collected, one at Time Point One (T1) and the other at Time Point Two (T2). Two sets of items parameters at the two time points (T1 and T2) were calibrated and scaled. Then a modified root mean square difference was used to quantify the differences between the two sets of item parameter estimates. After that, a simulation study was performed to establish the baseline of random variation. Finally, the differences between T1 and T2 items parameters were compared to the baseline to determine whether scale drift was observed.

Figure 1: Flowchart of the Design


31 studied items
T1 data collection, calibration & scaling
T2 data collection, calibration & scaling
Simulation with T1 examinees' theta & item parameters and calibration & scaling with T1 CAT operational responses
T1 item parameters
Baseline of random variation
T2 item parameters
Comparing RMSD of T1 and T2 TCCs against simulation baseline

Data Collection and Processing

A special data collection method was designed to get appropriate item parameter information for the study. A group of 31 GMAT® quantitative items were identified, divided into four groups and used as our studied items. Each group of items was administered online to a random sample of GMAT® test takers. These items were then calibrated and scaled together with operational items to collect data at T1. About 20 months later, the same items were administered under the same conditions to collect the T2 data. This time interval was chosen according to the testing cycles of the GMAT® program. In both administrations, the items were administered linearly, imbedded in the CAT test operational items and then calibrated in PARSCALE (The Educational Testing Service internal version of Muraki & Bock (1999)) using an item-specific-prior method (Folk & Golub-Smith, 1996; Guo, Stone & Cruz, 2001) to keep both calibrations on the same reference θ scale. If the group of studied items that were administered and calibrated in this manner showed no changes in their item parameters between T1 and T2, this would be taken as evidence of scale stability. In other words, this was like taking one snapshot of the reference θ scale at T1 and then another at T2, respectively. Any changes in the scale would be reflected in the item parameters derived at T2.

However, the item parameters from T1 and T2 will not be identical, even if the scale remains stable during the period between T1 and T2. This is because item parameters from the calibrations are only estimates of their true parameters and calibration errors are included in the estimates. In order to allow evaluation of scale stability, we assume that any differences in the item parameters between the two calibrations can include both random variations (calibration and scaling errors) and variations due to scale drift:

$$D = RV + SD \qquad (1)$$

where D, RV, and SD are for estimated difference, estimated random variations, and estimated scale drift, respectively.

With observed data, it was not possible to separate true changes in item parameters, or a scale drift, from random variations. A simulation study was designed and implemented to estimate the magnitude of random variations, which served as a baseline. If differences in the item parameters from T1 and T2 were smaller than the baseline, they were taken as random variations only and no scale drift was observed. If they were beyond the probable sizes of random variations, it could be concluded that scale drift was observed.

Simulation Study

This simulation was to establish an empirical baseline for random variations as a result of calibration and scaling processes over time. GMAT® quantitative CAT measure was used as our example. All the item parameters, examinee theta, and examinees' CAT responses were directly drawn from the GMAT® quantitative operational data. The simulation was conducted in the following steps:

1. From the T1 operational examinee records (N = 169,111), ten random samples of 1,000 examinees each were drawn without replacement.
2. For each sample drawn, the 1,000 examinees' θ values and the item parameters of the 31 studied items estimated from the T1 data were used to generate new item response vectors for the 1,000 examinees.
3. The item parameters of the 31 studies were calibrated for each sample using the 1,000 examinees' simulated item response vectors from step 2 together with the examinees' responses to the operational CAT items in the T1 administration. This concurrent calibration with the operational items and their item-specific priors put the new item parameters on the same reference θ scale as the operational items as well as the T1 and T2 studied items.

The rationale for the simulation is that, because the ability and item parameters from T1 were treated as true values and were used in generating item responses for item calibration in step 3, any differences in the calibration results from the 10 random samples can be attributed to random variations only. The results from the 10 simulated sample data can be used to establish an empirical baseline to estimate the degree of the random variation due to calibration and scaling processes. Ideally, a much larger number of random samples (say, 100 or more) could be used, which of course would be more time consuming and would require more resources. For this study of this new method, ten samples were more manageable and served the purpose of this study to explore the new method.

Analysis

A modified root mean squared difference (RMSD) on test characteristic curves (TCC) was developed to quantify the differences in estimated item parameters at T1, T2, and the ten simulation results. The RMSD between two TCCs can be expressed as follows:

$$RMSD_{TCC} = \sqrt{\frac{\sum_{\theta=-4}^{4}(\tau_{\theta,T2} - \tau_{\theta,T1})^2}{m}} \qquad (2)$$

where $\tau = \sum_{i=1}^{k} P_\theta$ for the k studied items, and m is the number of points on the θ ability scale.

In this study, m = 81 as all the items were evaluated at 81 points ranging from –4.0 to 4.0 in increments of 0.1. P_θ is an item's conditional probability of getting a correct answer, $P(X_i = 1 | \theta, a_i, b_i, c_i)$. P_θ is on a 0-to-1 metric, whereas τ is on a 0-to-k metric for k dichotomously scored items. However, because the $RMSD_{TCC}$ in Formula (2) is on the 0-to-k metric, it is not convenient to compare RMSD values among different tests that have a different number of items. Formula (2) can be modified to standardize the RMSD calculation by k, the number of items, such that the RMSD is now on the same 0-to-1 scale as P_θ. This is in effect an RMSD averaged over k items. In addition, population ability distribution can be included in calculating the TCC differences. Therefore, Formula (2) can be rewritten as:

$$RMSD_{TCC} = \sqrt{\frac{\sum_{\theta=-4}^{4} W_\theta (P^*{}_{\theta,T2} - P^*{}_{\theta,T1})^2}{81}} \qquad (3)$$

where $P^* = \tau / \mathrm{k}$, a proportion of correct value; W_θ is the population weight at θ.

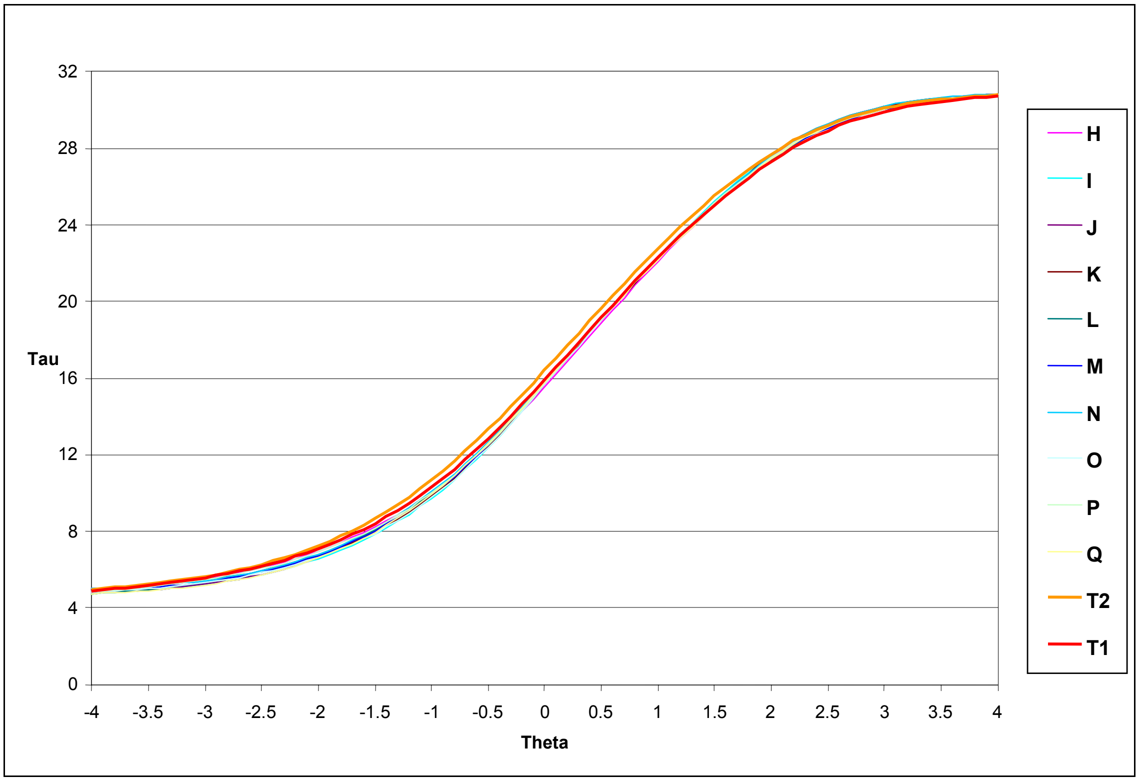
This is what was used for $RMSD_{TCC}$ statistics for TCC comparisons between T1 and T2, as well as T1 and the ten simulation results.

Results and Discussions

Differences in TCCs

Figure 2 displays the TCCs for T1, T2, and the ten simulations labeled H through Q. If there was any difference between T2 and T1 and the difference was really due to random variations from the calibration process, then the TCCs for T1 and T2 would be expected to be commingled with the ten simulated TCCs that reflect random fluctuations only. All the TCCs appear to be very close to one another with no one distinctly departing from the rest. This observation suggests that, if there were any differences between T2 and T1, the differences were very small and the differences could mainly be due to random variations.

Figure 2: TCCs for the T1, T2 and the Ten Simulated Results



Upon closer examination of the TCCs, the T2 line appears to be above the other TCC lines across most part of the ability scale (roughly from –3 to 2.0, see Figures 3, 4, and 5). At the 2.0 and higher part of the ability scale, the T2 line merges into the other lines. That the T2 line is consistently higher than the T1 line and the ten simulated lines over most of the scale does warrant a tentative conclusion that there was a difference between the T2 and the T1 lines and the difference could not be attributed to random variations only. Of course, this conclusion was based on visual judgments only and is not enough for the purpose of this study. The modified RMSD method was used to compare the statistics of the differences to determine if the differences would be meaningful.

Figure 3: TCCs at the Lower Range of the Theta Scale

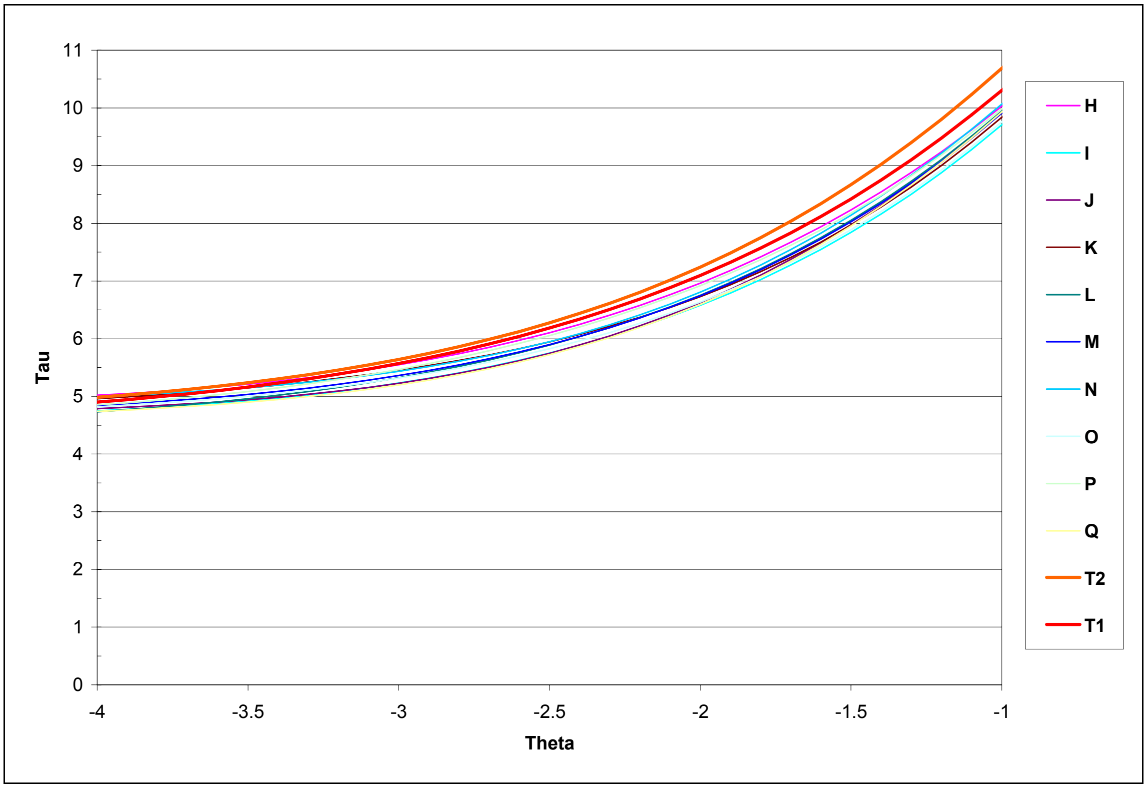


Figure 4: TCCs at the Middle Range of the Theta Scale

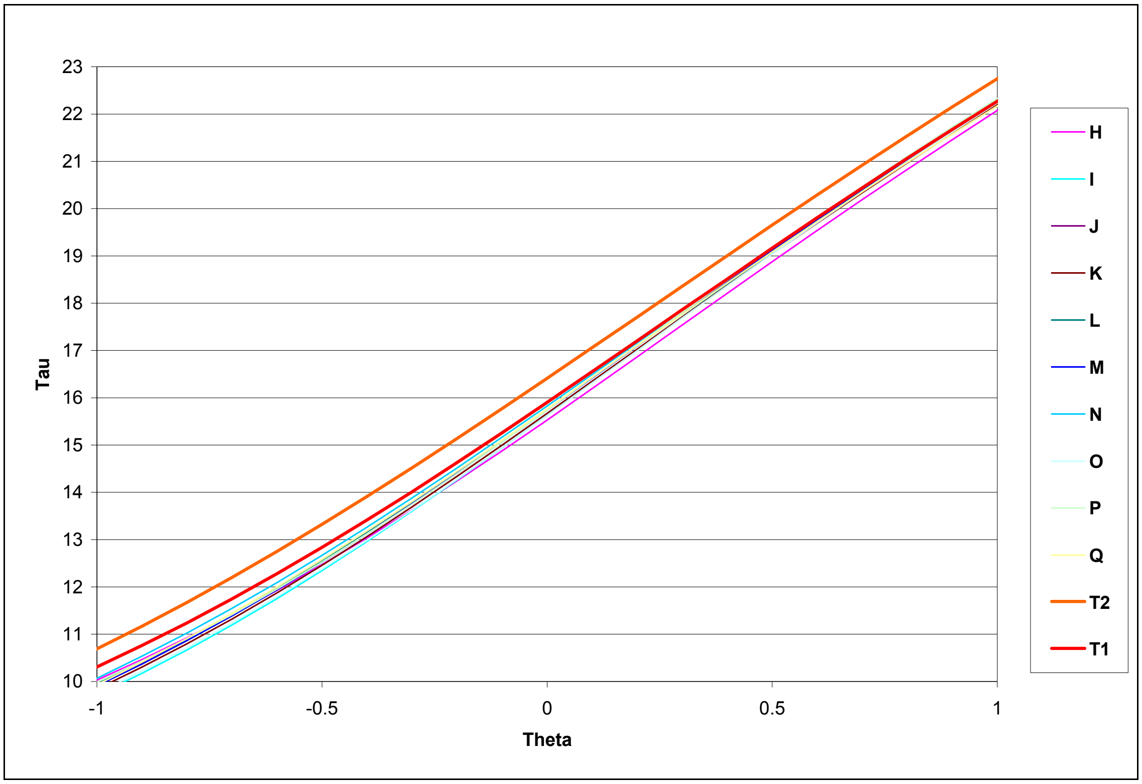
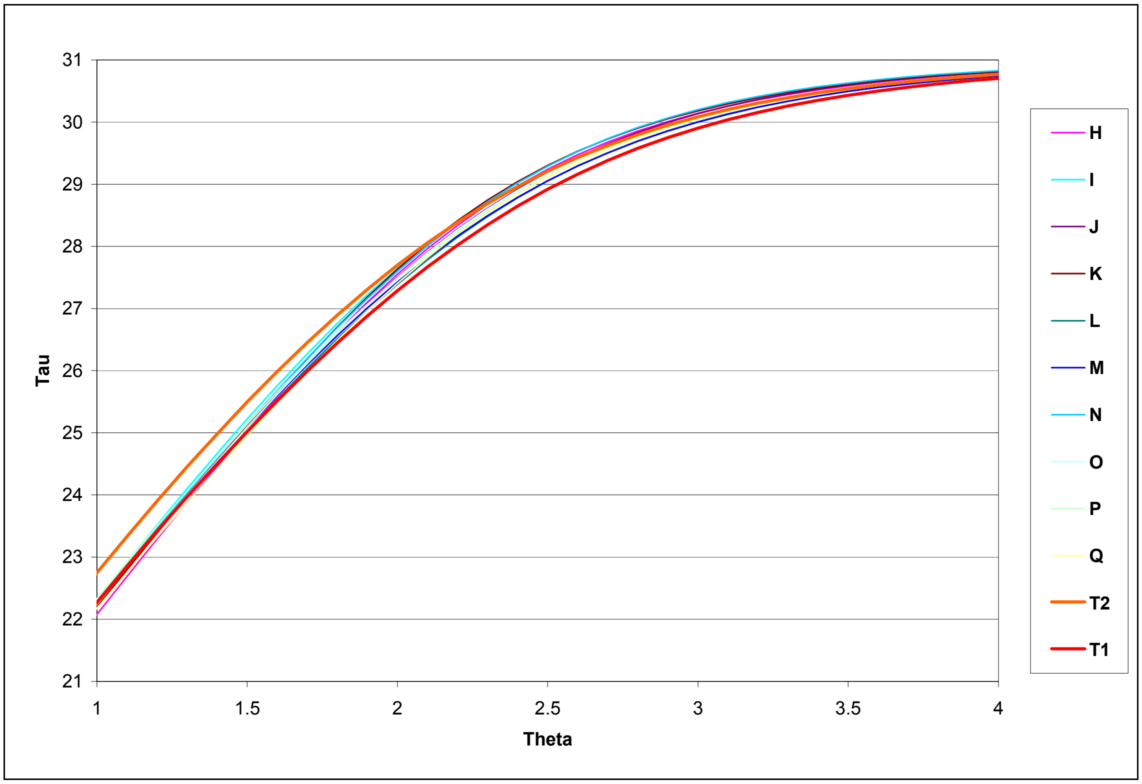


Figure 5: TCCs at the Upper Range of the Theta Scale



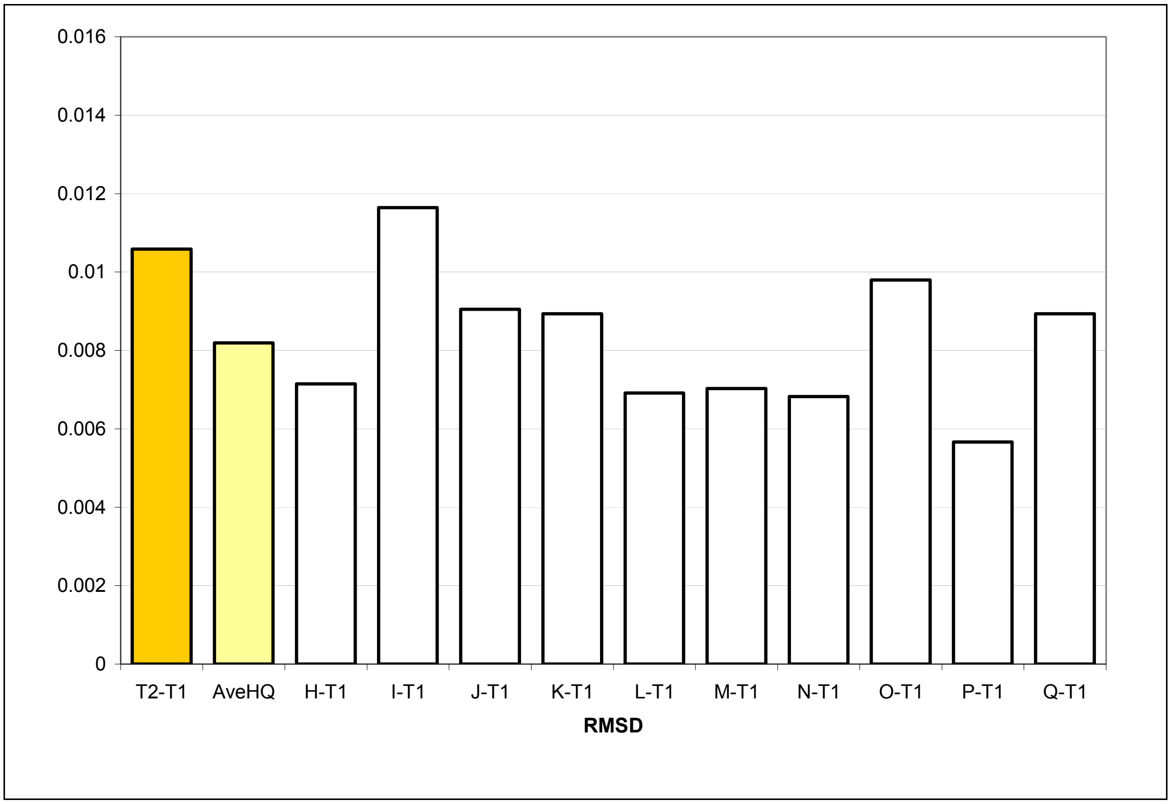
Additionally, the T1 line is found to be slightly above the ten simulated lines but below the T2 lines mainly between –3 to 0 on the ability scale. The difference between the T1 line and the ten simulated lines also appears to change directions with a slight clockwise shift. The T1 line is above the 10 simulated lines in the lower to middle part of the scale and shifts below the ten lines at the top part of the scale.

Modified RMSD Results

The modified RMSD statistics were calculated using Formula (3) for the TCC difference between T2 and T1 and for the differences between each simulated TCC (H through Q) and the TCC for T1. Both unweighted and weighted RMSD can be computed, depending on whether population ability distributions need to be taken into consideration. With the unweighted RMSDs, the W_θ in Formula (3) was set to 1 for all θ points. In this study, only the unweighted RMSD method was used.

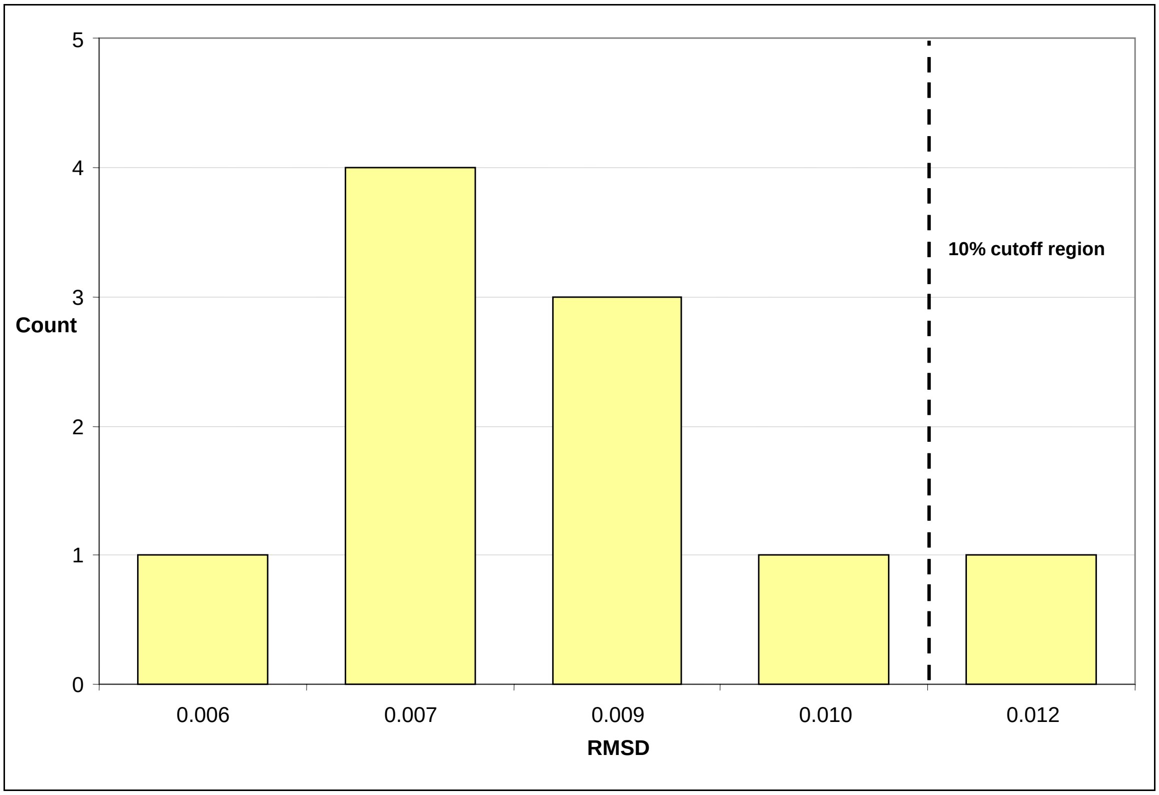
The unweighted RMSDs for the TCC comparisons are depicted in Figure 6. The left-most column is RMSD value for T2 minus T1 (T2 – T1). Next to T2 – T1 is the average RMSD (AveHQ) for the 10 simulations minus T1. The RMSDs for the 10 individual simulations minus T1 (H – T1 to Q – T1) are also shown in the figure. No particular pattern is observed in the differences across the ten pairs from H – T1 to Q – T1. The average of the 10 RMSDs is 0.00819 and the range of the 10 RMSDs covers from 0.006 (P – T1) to 0.012 (I – T1). The RMSD for T2 – T1 was 0.011 and was within the range of the 10 RMSDs that are used as the baseline.

Figure 6: RMSD Statistics for T2 Minus T1 and the Ten Simulations Minus T1



While Figure 6 reveals some distributional characteristics of the RMSD in comparison with the T2 – T1 RMSD, the focus of interest, such information does not provide a means of assessing if the T2 – T1 RMSD can be considered within random variation or if it indicates some degrees of scale drift. To assess if an observed RMSD like that of T2 – T1 is due to random variation only or also due to scale shift, a cutoff point on the distribution of the RMSDs needs to be established. This distribution can be an empirical one since no theoretical one is available. This empirical distribution is the distribution of the RMSDs for the simulation results minus T1. In this study, it's the distribution of the ten RMSDs. Figure 7 displays the ten RMSD values that were rounded to the third decimal place and the number of simulation results by the rounded RMSD value points. If a large number of simulations (100 or more) are run, the resulting distribution of the RMSD values will probably cover the whole RMSD scale and show a certain shape (maybe something resembling a χ^2 distribution).

Figure 7: Empirical Distribution of the Ten RMSD Values



Regardless of the shape of the curve, a cut-off point can be selected as the criterion for evaluating an RMSD like that of T2 – T1. For example, a cut-off point can be set at the 95th or 90th percentile on the RMSD scale and a T2 – T1 RMSD is declared exhibiting evidence of scale drift if it falls beyond the cut-off point. Stated another way, the probability is around 0.05 or 0.1 for an RMSD value to be greater than the cut-off value when that RMSD is still due to random variation. This is the same logic as in statistical hypothesis testing. The application here may be expressed as follows:

H_0: T2 – T1 RMSD = random variation

H_a: T2 – T1 RMSD > random variation.

The only difference here from a formal statistical hypothesis testing is that no test statistic from a known theoretical distribution is used. Instead, the 95th or 90th percentile cut-off point is associated with the distribution of the RMSD values from simulation results.

The distribution in Figure 7 can be used for this purpose even though there are only ten RMSD values. With the ten values, it is not possible to set a cut-off at the 95th percentile. Therefore the 90th percentile point is used as the cut-off for this example. The P – T1 RMSD is 0.012, the highest of the ten, and is the cut-off point below which an RMSD is considered as being within random variation. This is analogous to doing statistical hypothesis testing using an alpha level of 0.1. If a T2 – T1 RMSD falls in the 10% cut-off region, the H_0 is rejected and this RMSD is declared exhibiting scale drift. On the other hand, if an RMSD falls below the cut-off point, the H_0 is retained and the RMSD is considered due to random variation only. Since the T2 – T1 RMSD of 0.011 is smaller than the 0.012 cut-off RMSD value, the difference between T2 and T1 can be considered due to random variation in calibration and no evidence of scale drift in the GMAT® Quantitative measure is observed between T1 and T2.

Conclusions

Scale stability is an important quality of a large-scale CAT program and should be maintained through research on scale drift evaluations in the CAT operations. Because any differences in the parameter estimates for the same items over time include both random variations due to calibration and systematic change due to scale drift, it is necessary to be able to disentangle the scale drift component of the total change from the random variation component. In this study, a special online data collection method was implemented and a modified root mean squared difference statistic was used to measure the difference in item parameters between the two time points. Then an empirical baseline was established using simulations for evaluating the difference. The result showed that scale drift was not detected in the GMAT® Quantitative measure and the observed differences between the two sets of item parameters calibrated at two time points were random variations.

Because there is scarcely anything in the CAT literature on evaluating scale drift using both observed and simulated data, methods used in this study will make it possible for researchers to perform scale drift studies. For those who may want to use these methods, it is helpful to point out that this study focused on the methods themselves and that only ten simulations were conducted and each used only about 1,000 examinees. For a real evaluation of the scale drift in a CAT program, we have the following two recommendations. First, the number of simulation runs should be at least 100 in order to obtain an adequate sample of the simulation results for evaluation. Second, the simulation parameters should be the same as those in real pretest calibrations, such as the IRT model, sample size, calibration and scaling methods. This would yield more stable and realistic estimates of the parameters.

Authors

Fanmin Guo is the Director of Psychometric Research at the Graduate Management Admission Council®, McLean, VA. Lin Wang is a Lead Measurement Statistician in charge of the English language assessment programs in Center for Statistical Analysis, Educational Testing Service, Princeton, NJ.

Acknowledgements

Earlier versions of this paper were presented at the 2003 Annual Meeting of the American Educational Research Association (AERA) in Chicago, USA and, with an improvement on the method, at the 2005 AERA Annual Meeting in Montreal, Canada.

Contact Information

For questions or comments regarding study findings, methodology or data, please contact the GMAC® Research and Development department at research@gmac.com.

Reference

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service.

Glas, C. A. W. (1998). Quality control of on-line calibration in computerized assessment. (Research Report 98-03). The Netherlands: University of Twente.

Folk, V., & Golub-Smith, M. (1996, April). Calibration of on-line pretest data using BILOG. Paper presented at the annual meeting of NCME, Chicago.

Guo, F., Stone, E., & Cruz, D. (2001, April). On-line calibration using PARSCALE item specific prior method: Changing test population and sample size. Paper presented at NCME Annual Meeting, Seattle, Washington.

Guo, F., & Wang, L. (2003, April). Online calibration and scale stability of a CAT program. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL

Kolen, M. J. (1999-2000). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment, 6*(2), 73-96.

Kolen, M. J., & Brennan, R. L. (1995). Test equating methods and practices. New York: Springer-Verlag.

Muraki, E., & Bock R. (1999). PARSCALE 3.5: IRT item analysis and test scoring for rating-scale data. Scientific Software, Inc.

Stocking, M. (1988). Scale drift on-line calibration. (ETS Research Report 88-28-ONR). Princeton, NJ: Educational Testing Service.

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement*, 38(1), 19-49.