Test Security for Multistage Tests: A Quality Control Perspective

Charles Lewis, Yi-Hsuan Lee, and Alina A. von Davier

Educational Testing Service

Paper presented at the Conference on Statistical Detection of Potential Test Fraud

May 23-24, 2012, Lawrence, Kansas

Abstract

A multistage test (MST) is a computer-based assessment that may be thought of as a compromise between a linear test and a computer-adaptive test (CAT). As such, MSTs may be vulnerable to at least some of the major security threats associated with each of these types of test (e.g., copying for linear tests and item pre-knowledge for CATs). The degree of vulnerability of any particular MST to these threats, as well as others, will depend (among other things) on details of the MST assembly and administration design. To supplement these preventative measures, routine statistical monitoring of response and timing data for items, modules and tests, as well as the screening of performance of individual test takers and clusters of test takers, is essential. We strongly believe that test security procedures are properly understood in terms of quality control for a testing program, and that the goal of these procedures should be for the program to report only valid test scores, while treating all test takers fairly.

After a general introduction to issues of test security for MSTs, two examples of monitoring procedures are discussed, one for inconsistent test taker performance and one for inconsistent item performance.

Test Security for Multistage Tests: A Quality Control Perspective

A multistage test (MST) is an assessment that differs from a both a computer-adaptive test (CAT) and a linear test in several respects, and may be thought of as a compromise between these two testing formats. An MST adopts a multistage adaptive design that provides adaptation across test takers at the section or module level rather than at the item level. Figure 1 shows an example of an MST design with two stages and a total of four modules. Items in each module are delivered in a fixed order. All test takers receive the routing module, which typically contains items with a broad range of difficulties. According to their performance on the routing module, the test takers are routed to one of the second-stage modules, with items of higher (Hard module), medium (Medium module) or lower difficulties (Easy module). The routing decision is made based on pre-determined cut scores.

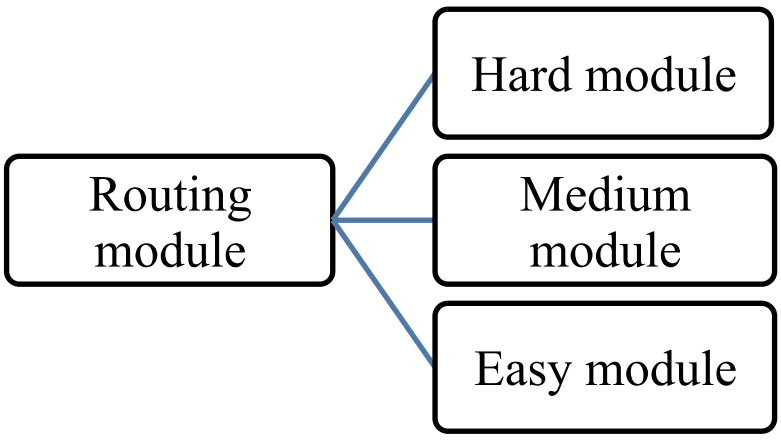

Figure 1. A hypothetical two-stage MST

Quality control procedures must be part of any testing program, including those using MSTs. This starts with design considerations to minimize a variety of threats to the validity of reported test scores. However, there also needs to be routine statistical monitoring of response and timing data for items, modules and tests, as well as the screening of the performance of individual test takers, clusters of test takers, and test centers. Items that are not functioning as expected should be identified as rapidly as possible, and test takers whose performance is questionable should not have their scores reported to test users.

A wide range of standard test security procedures has been developed for CATs and for linear tests. Many of these procedures may also be appropriate for use with MSTs, depending on how the test is designed, assembled and administered. For example, if a single routing module is administered to two test takers located next to each other in a test center, then an opportunity for copying or communication exists, just as when a single linear form is administered. In this case, standard procedures used to detect unusual similarity among responses of two or more test takers may also be applied to MST responses.

To take another example, if the modules to be administered for an MST are taken from a limited pool over an extended time period, then item pre-knowledge would be a threat, just as it is for a CAT. In this case, procedures used with CATs to detect inconsistent responses or unusual response latencies may also be appropriate for MSTs. The same point applies to the monitoring of item performance. It is possible to use methods developed for CATs and linear tests to look for unusual distributions of

responses or response latencies for MST items. If an item is used in a single module over an extended period of time, or is used in multiple modules, then statistical process control methods may be appropriate for tracking its performance.

In the sections that follow, two illustrations of quality control procedures that may be appropriate for MSTs are given. First, an example is given of a comparison of test taker scores on operational MST modules and scores on a pre-test module, looking for inconsistent performance. Such a comparison could form the basis of a routine screening procedure if there is a concern about possible item pre-knowledge, as well as possible communication or copying among test takers. Second, a simulation is described, looking at the application of a standard statistical process control method (Cusum) to the monitoring of item performance.

Analysis to Detect Inconsistent Variable Module Performance

A group of test takers are identified, who have been administered the same sequence of operational modules and the same variable module (typically containing pretest items). A linear regression of the number of correct responses for the variable module on the total number of correct responses for the operational modules is obtained. Test takers are identified with an unusually low number of correct responses on the variable module, compared to the predicted number of correct responses based on their operational score. This information could be used in combination with other sources of information as a basis for questioning these test takers' operational scores. The reasoning behind this analysis is that, for a variety of reasons, test takers may be more likely to have access to answers to operational items than to pretest items contained in a variable module. (See, for instance, Haberman, 2008.)

In the present example, *N*=1,060 test takers were found with responses to the items in the same set of operational and variable modules. Depending on the testing volume relative to the number of modules in use at any one time, this sample size might easily be substantially greater or substantially smaller. In the latter case, there would be reasonable concern about sampling error, and Bayesian methods might be useful. In the present case, simple analyses are sufficient for the purposes of quality control. Figure 2 shows the distribution of the total number-correct scores for the operational modules. The relatively short tails of the frequency distribution are not surprising in the context of MST.

For the same group, Figure 3 shows the distribution of the number-correct scores for a variable (non-operational) module. In contrast to the distribution of the operational scores, this distribution is highly skewed, suggesting that the items in this module were too easy for this group. Such a distribution might also be expected to present a challenge for obtaining useful results from a standard linear regression analysis.

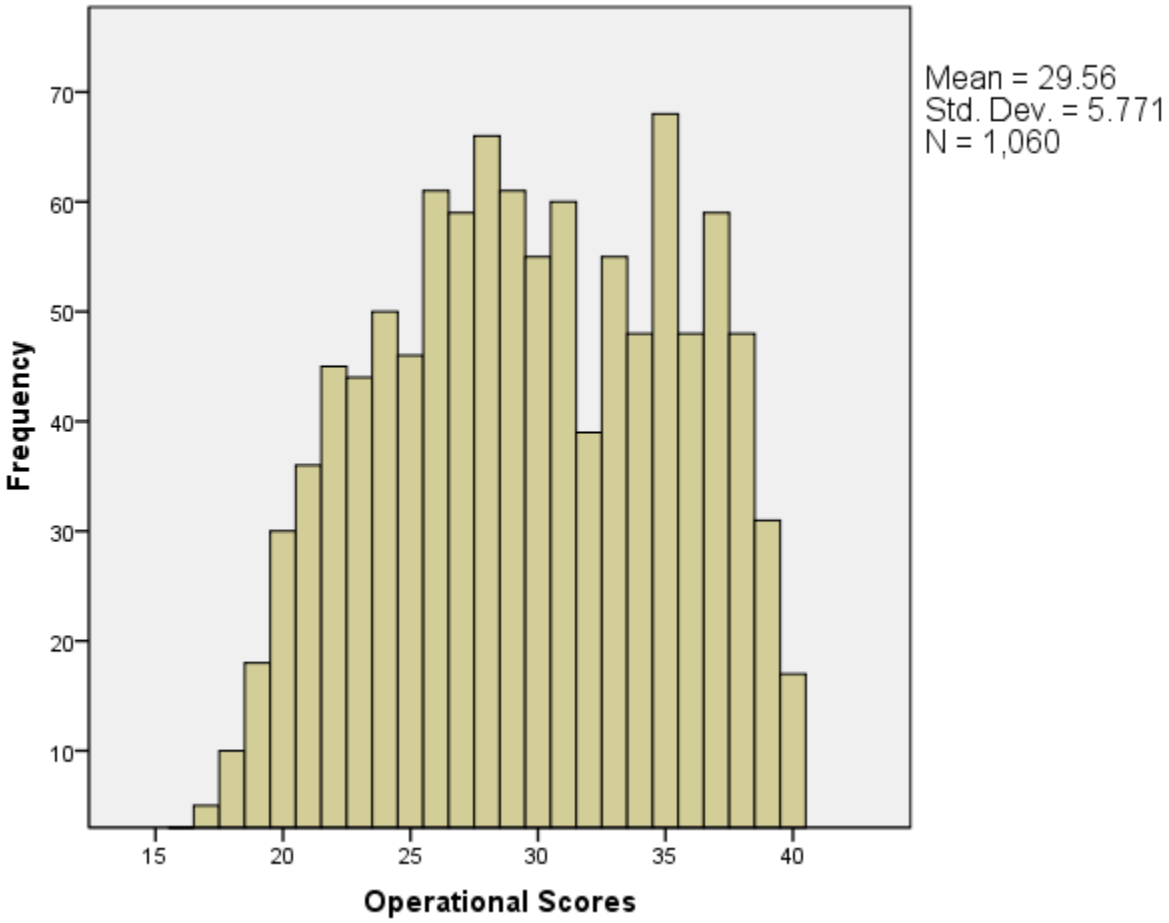

Figure 2. Frequency distribution of operational number-correct MST scores for selected group

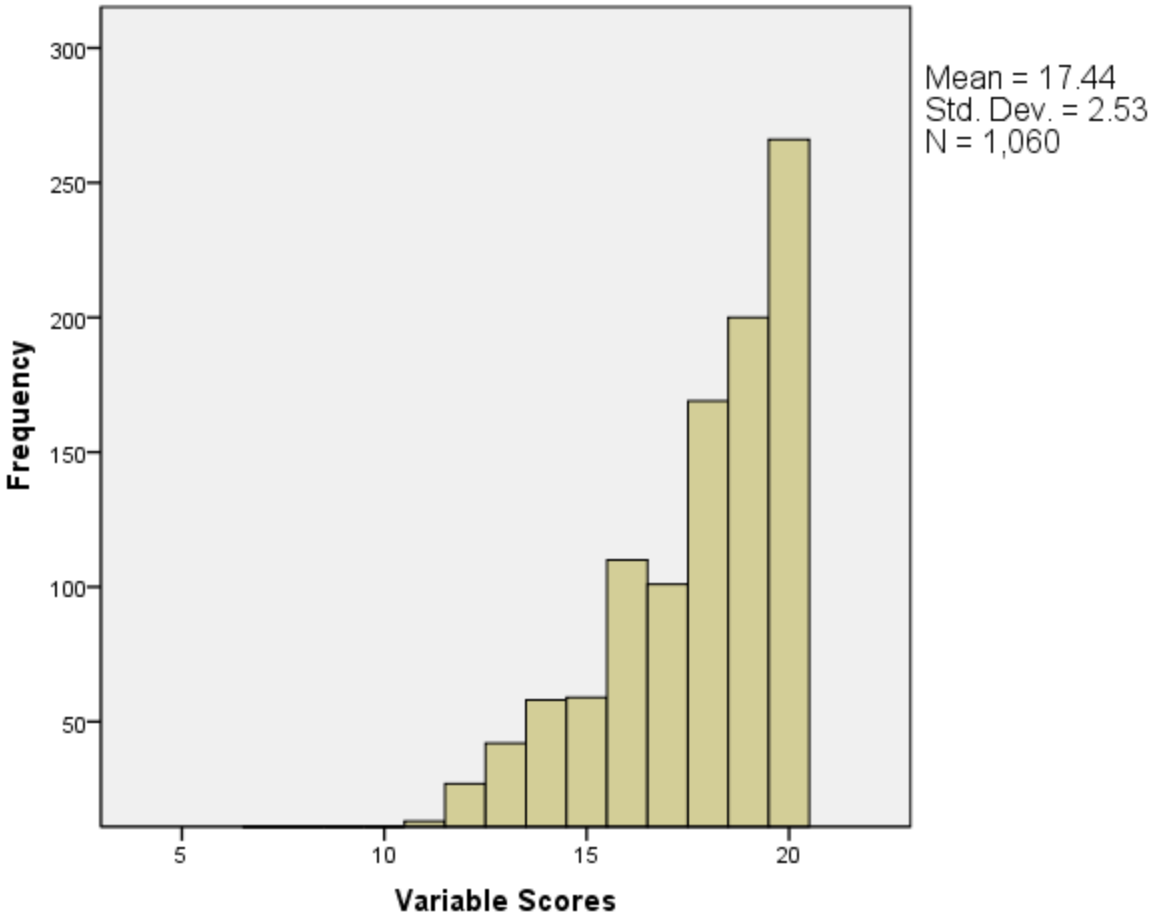

Figure 3. Frequency distribution of variable (non-operational) number-correct MST scores for selected group

Table 1. All cases with standardized residuals exceeding 3.0 in absolute value

Case Number	Std. Residual	Operational Score	Variable Score	Predicted Value	Residual
1	-3.990	22	8	15.07	-7.072
2	-3.671	17	7	13.51	-6.506
3	-3.460	19	8	14.13	-6.132
4	-3.358	28	11	16.95	-5.950
5	-3.358	28	11	16.95	-5.950
6	-3.215	24	10	15.70	-5.698
7	-3.181	27	11	16.64	-5.637
8	-3.289	34	13	18.83	-5.829
9	-3.249	21	9	14.76	-5.759
10	-3.073	20	9	14.45	-5.445
11	-3.073	20	9	14.45	-5.445
12	-3.073	20	9	14.45	-5.445
13	-3.147	30	12	17.58	-5.577
14	-3.073	20	9	14.45	-5.445
1060	3.311	19	20	14.13	5.868

Note. $N = 1{,}060$.

The Pearson product-moment correlation between the operational and variable scores for this sample is $r = 0.714$. The corresponding linear least squares regression of variable scores on operational scores is given by

$\hat{y} = 8.183 + 0.313x$.

The estimated standard error of the residuals from this regression is $\hat{\sigma}_e = 1.772$. The minimum and maximum of the residuals are -7.072 and 5.868, respectively. Table 1 provides a list of cases with extreme residuals for this sample. Figure 4 shows the

distribution of the standardized residuals. This distribution has a slight negative skew and somewhat heavy tails, relative to a Normal distribution.

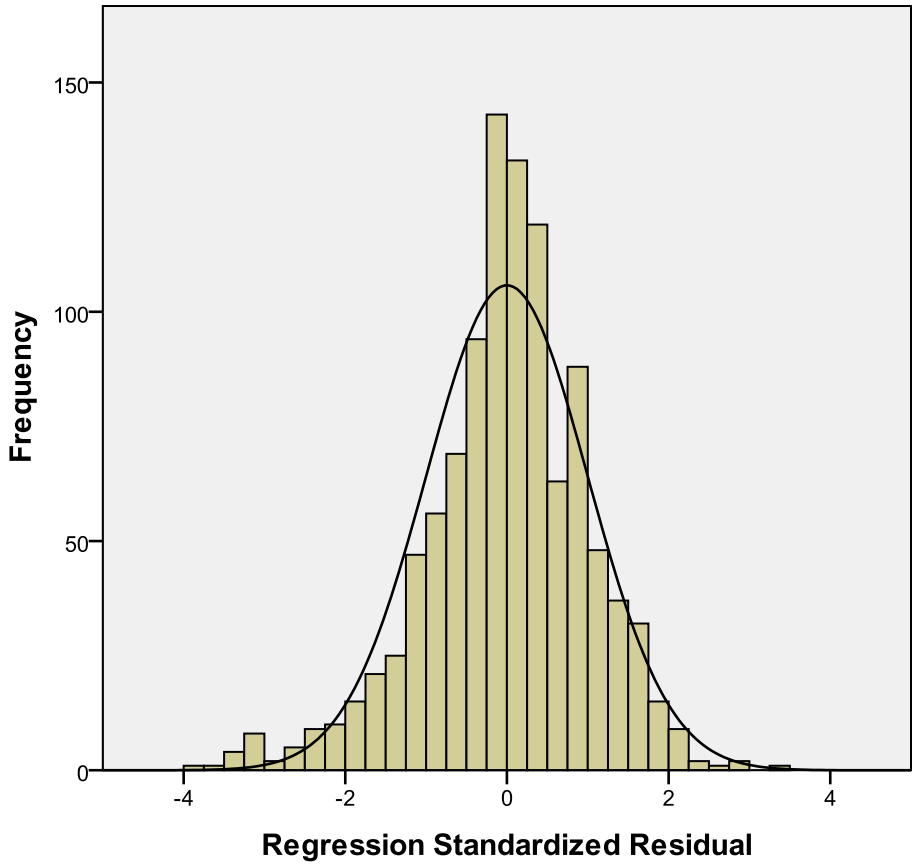

Figure 4. Frequency distribution of standardized residuals from linear regression for selected group

For purposes of illustration, consider case number 1 in Table 1. This test taker has an operational score of 22 and a variable score of 8. The predicted variable score, based on the operational score, is 15, so the residual is -7, and the standardized residual is -4. Figure 5 represents the conditional distribution of variable scores using a Normal density with mean equal to 15 and standard deviation equal to 1.8, together with an

indication of the test taker's actual variable score of 8, making clear its status as an outlier.

It is important to note that such a result taken alone would not normally be considered sufficient to cancel this test taker's score. Instead, it might lead to additional examination of aspects of the test taker's performance as well as other considerations that, taken together, might provide justification for such an action.

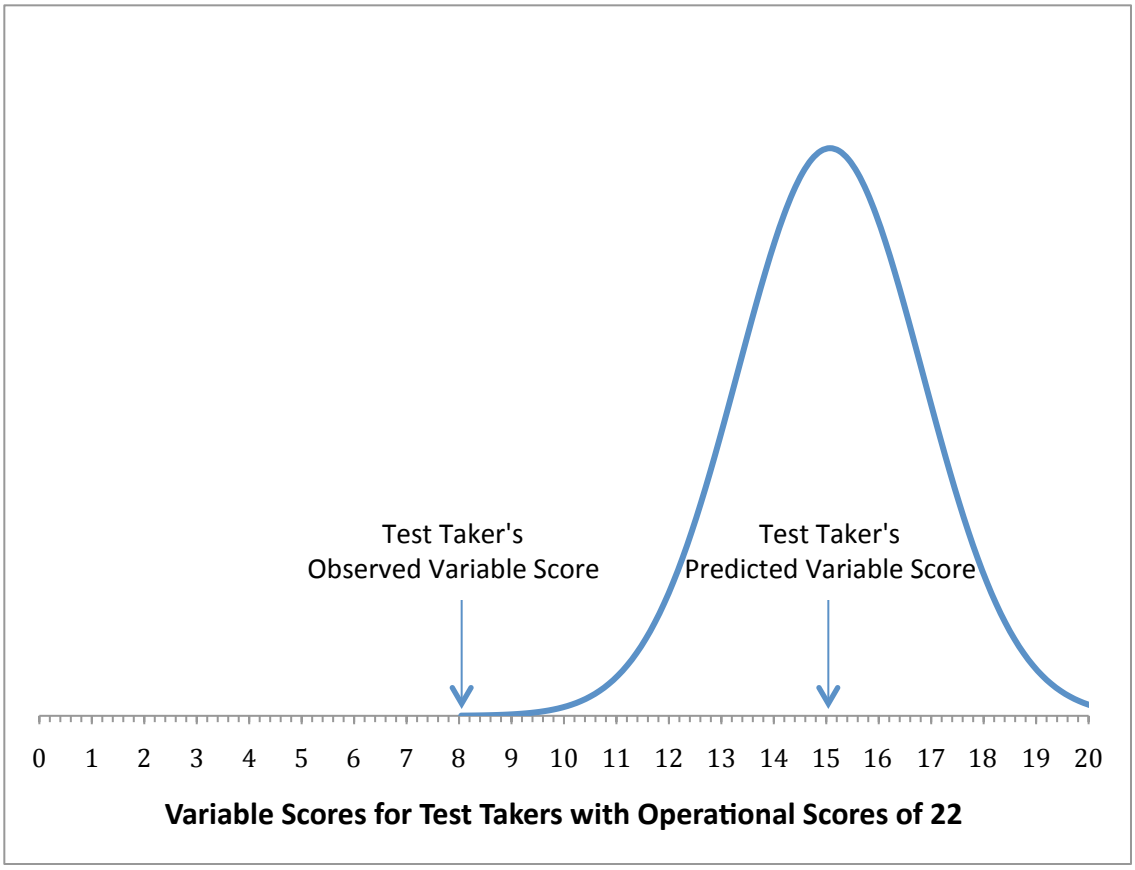

Figure 5. The theoretical conditional distribution of variable scores, given an operational score of 22, showing the test taker with a variable score of 8 as an outlier

Simulation of an Analysis to Monitor Item Performance for an MST

The primary motivation for monitoring item performance for an MST is the assumption that its items will be re-used. This is a state of affairs that MST shares with many other types of tests, including but not restricted to most computer delivered tests. MST item re-use could occur because a fixed set of modules is used for an extended period of time (such as a week or a month). It could also occur when modules are only used for a short period of time (such as an hour or a day), after which their items are returned to an item bank, where they become available for use in other modules at some point in the future. In any event, it is important to know if items being considered for re-use are performing differently than when they were first calibrated. This state of affairs might lead to re-calibration or removal from further use.

The Cusum procedure (see, for instance, Montgomery, 2008) is a standard method for monitoring the quality of a statistical process over time. In the version employed here, a statistic z_i is computed for case *i*, with $i = 1,2,\cdots$. When the process being monitored is in control, each value of this statistic is assumed to be an independent random sample from a standard Normal distribution. If the process is out of control, this is assumed to result in an *increase* in the mean value of z_i. The actual Cusum procedure uses two positive constants, *k* and *h*, that must be specified by the user and requires the definition of a sequence based on z_i as follows:

Let $S_0 = 0$.

For $i = 1,2,\cdots$, let $S_i = \max\left\{0, S_{i-1} + z_i - k\right\}$.

The procedure stops for the first case with $S_i > h$.

The values of the constants *k* and *h* control properties of the procedure (analogous to Type I error rate and power for standard hypothesis testing). Commonly chosen values are $j = 0.5$ and $h = 5.0$. With these choices,

$S_i = \max\left\{0, S_{i-1} + z_i - 0.5\right\}$, and the procedure stops for the first case with $S_i > 5.0$.

In the simulation reported here, it is assumed that the probability of a positive item response $\left(y_i = 1\right)$ is given by the 2PL IRT model:

$$\Pr\left(y_i = 1 \middle| \theta_i\right) = p_i = \frac{\exp\left[1.7a\left(\theta_i - b\right)\right]}{1 + \exp\left[1.7a\left(\theta_i - b\right)\right]}.$$

Assuming, for purposes of the simulation, that *a*, *b*, and θ_i are known, then z_i may be defined as

$$z_i = \frac{y_i - p_i}{\sqrt{p_i\left(1 - p_i\right)}}.$$

This definition is used when the item responses are analyzed one at a time. Of course, in this case, z_i does not have a standard Normal distribution when the process is in control (i.e., when test takers are responding to the item according to the 2PL model with the specified parameters): It does have a mean of zero and a standard deviation of one, but it is, nonetheless, a binary variable taking on the values

$z_i = -\sqrt{\frac{p_i}{1 - p_i}}$ when $y_i = 0$ and $z_i = \sqrt{\frac{1 - p_i}{p_i}}$ when $y_i = 1$.

One way to address this serious violation of a standard Cusum assumption is to work with groups of item responses. As a practical matter, the responses may only become available in groups (at the end of a testing session, for instance). If group *i* has n_i responses, labeled y_{ij} for $j = 1, \cdots, n_i$, with corresponding latent trait values θ_{ij} and item response probabilities p_{ij}, then the definition of z_i is modified to

$$z_i = \frac{\sum_{j=1}^{n_i} y_{ij} - \sum_{j=1}^{n_i} p_{ij}}{\sqrt{\sum_{j=1}^{n_i} \left[p_{ij} \left(1 - p_{ij} \right) \right]}}.$$

When the process is in control, z_i will have a mean of zero and a standard deviation of one. In many situations, it will be appropriate to approximate its probability distribution using a Normal density.

In the simulations of this process that were carried out for this study, the number of test takers per group was varied, as were the values of the item parameters. The values of θ_{ij} were independently sampled from a standard Normal distribution, and, in the simulation of the process in control, the item responses y_{ij} were then independently sampled from Bernoulli distributions with probabilities of correct responses given by the model values p_{ij}.

By way of illustration, simulation results are presented for an item with $a = 1.0$ and $b = 0.0$. Grouping was used to compute z_i, with $n_i = 10$. The process was simulated for a maximum of 5,000 groups, and 3,000 replications were carried out.

For the null case (where the process is in control and responses follow the 2PL model), Figure 6 shows the frequency distribution of "run lengths," expressed in terms of the first group number $\left(i_{crit}\right)$ for which $S_i > 5.0$, multiplied by $n_i = 10$ to give the total number of test takers whose responses have been analyzed at that point. The most notable feature of this distribution is its extreme positive skewness. The median of the distribution is 6,390. This may be interpreted as follows: For an item that is in control, the estimated probability is 0.50 that the item will be incorrectly identified (flagged) as being out of control by the time that 6,390 test taker responses (or 639 groups of responses) have been analyzed. The 10^{th} percentile of the distribution is 1,040, so the estimated probability is 0.10 that the item will be flagged by the time 1,040 responses have been analyzed. This null distribution characterizes the attrition rate, or rate of loss of items that are in control, as a result of employing the Cusum procedure.

To simulate an out of control process, the y_{ij} were independently sampled from Bernoulli distributions with probabilities of correct responses given by

$\tilde{p}_{ij} = c_i + \left(1 - c_i\right) p_{ij}$, with

$c_i = 0.0$ for $i = 1, \cdots, 100$ (null case),

$c_i = 0.1$ for $i = 101, \cdots, 200$ (initial exposure level),

$c_i = 0.2$ for $i = 201, \cdots, 300$, etc., and

$c_i = 0.8$ for $i = 801, \cdots$ (maximum exposure level).

This choice was intended to model the gradual exposure of an item with repeated use.

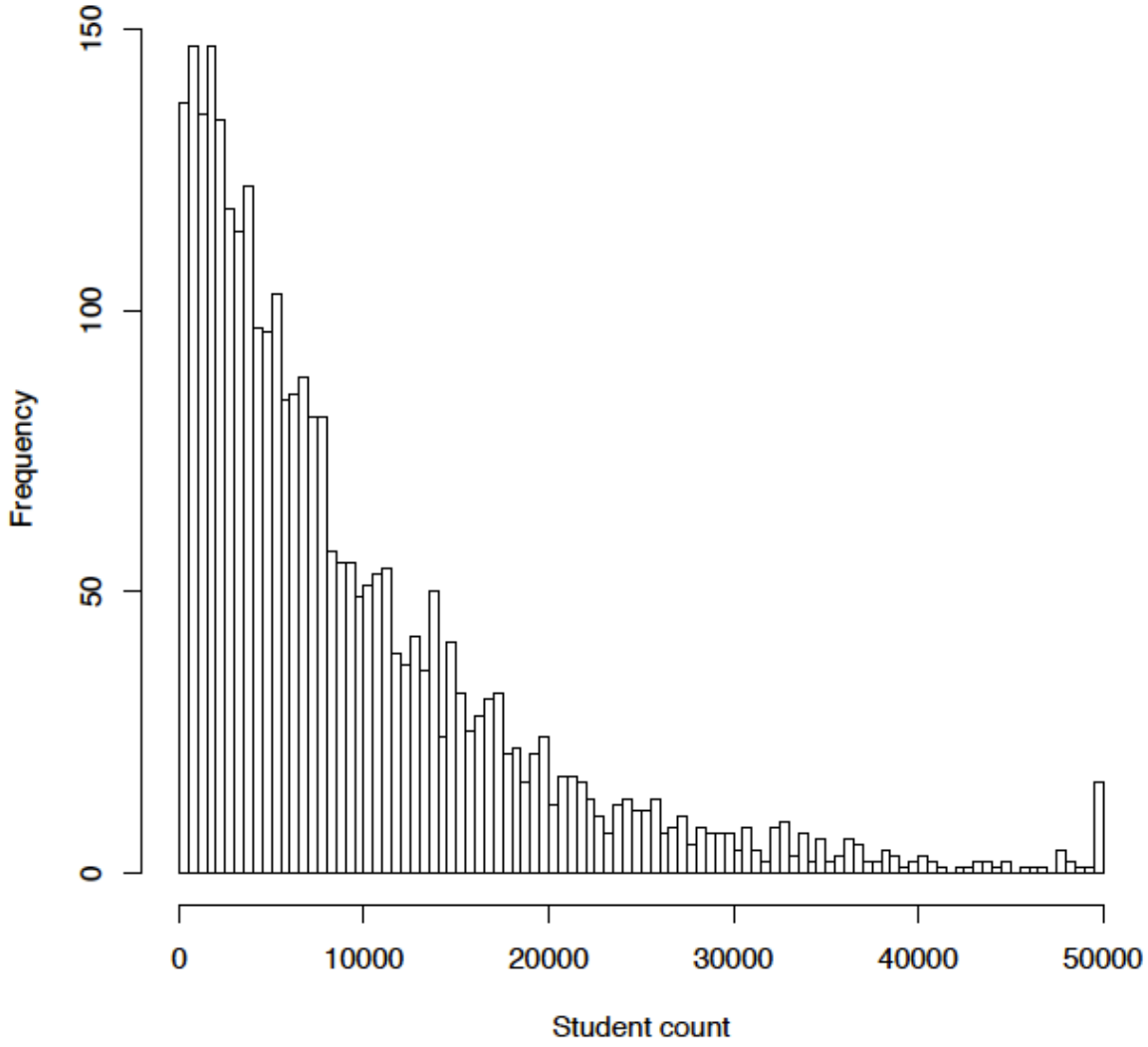

Figure 6. Frequency distribution when the process is in control (null case), of the number of test takers whose responses have been analyzed when the Cusum procedure falsely identifies the item as out of control. Responses were analyzed for groups of $n_i = 10$ test takers. There are a total of 3,000 replications.

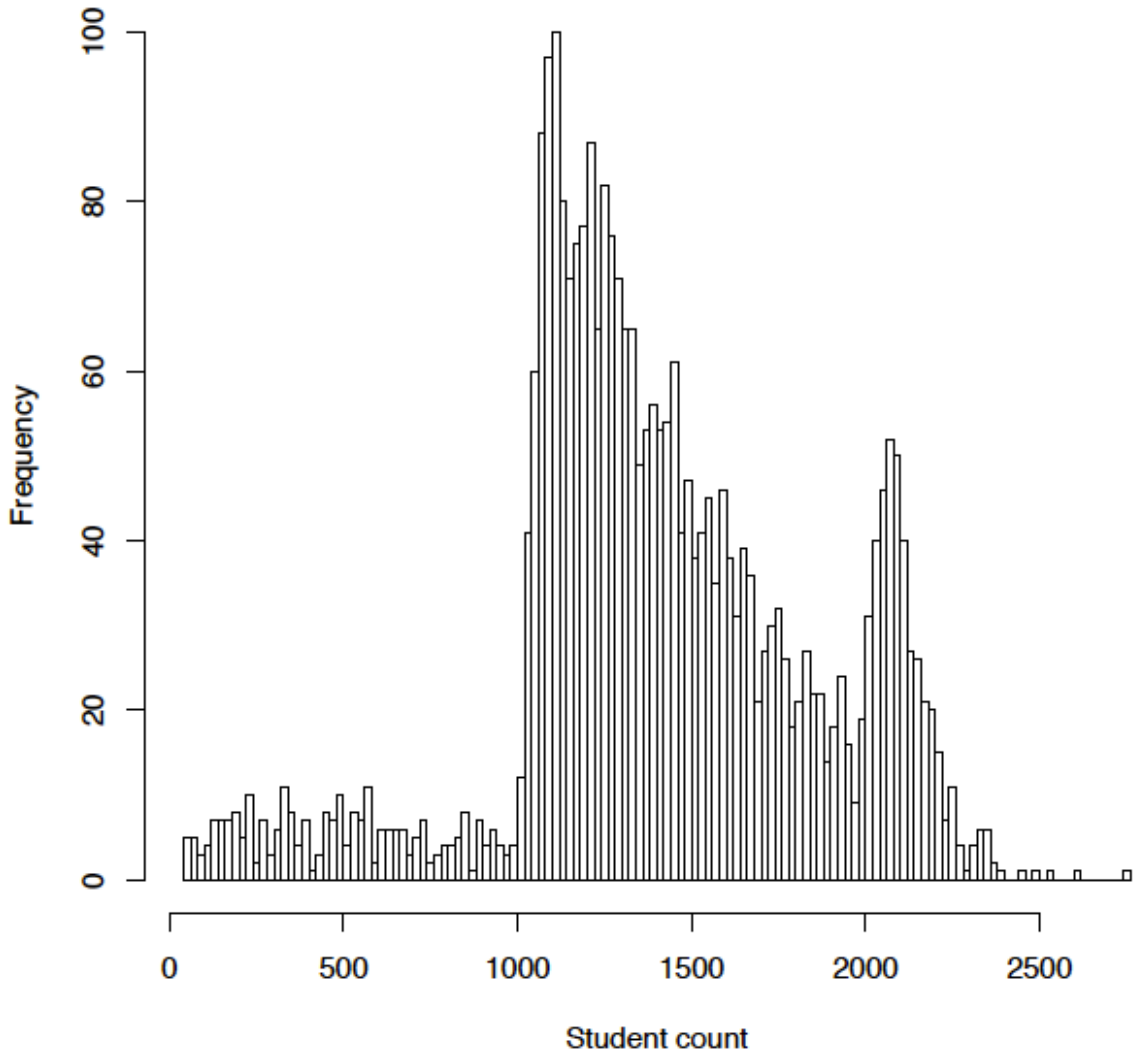

Figure 7. Frequency distribution when the process is out of control (gradual exposure case), of the number of test takers whose responses have been analyzed when the Cusum procedure identifies the item as out of control (falsely for the first 1,000 test takers and correctly thereafter). Responses were analyzed for groups of $n_i = 10$ test takers. There are a total of 3,000 replications.

Figure 7 shows the distribution of run lengths for this out of control case. The most striking feature of this distribution is its saw-tooth form, with the primary mode occurring after 1,060 responses have been analyzed. Recall that, starting with response number 1,001, the response probabilities were inflated using $c_1 = 0.1$. Also note that there is a secondary mode after 2,040 responses have been analyzed. (Starting with

response number 2,001, the response probabilities were further inflated using $c_2 = 0.2$.)

Virtually all of the 3,000 replications resulted in flagging of the item by the time 2,500 responses have been analyzed. This distribution's 10th percentile is 1,040, identical to that of the null distribution. In other words, just under 10 percent of the replications in this condition resulted in false positive flagging of the item while it was still in control. The median of this distribution is 1,360, so just over 40 percent of the replications resulted in flagging of the item after 360 potentially out of control responses were analyzed. This means that there is about a 40% chance that the item would be correctly flagged after approximately 36 aberrant responses were analyzed. Finally, the 90th percentile of the distribution is 2,060. This means that there is about a 80% chance that the item would be flagged after 1,060 potentially out of control responses were analyzed, of which an estimated 112 responses were actually aberrant.

The usefulness of this Cusum procedure for a given testing program would depend on an evaluation of the trade-off between the attrition rate for items in control and the detection speed for items out of control. In the present example, the question may be stated as:

> Is it acceptable to lose 10% of the items that are in control after about 1,000 exposures (responses) in return for being able to detect about 40% of the items that are out of control based on an average of 36 aberrant responses, and to detect about 80% of these items based on an average of 112 aberrant responses?

Finally, some limitations of the simulation should be mentioned. First, it was assumed that the in-control responses to the item followed the 2PL IRT model, and that the item parameters were known. A more realistic simulation would use estimated item parameters and allow non-systematic deviations of the response distribution from the 2PL model.

Second, it was assumed that the latent trait value was known. This assumption is (even) more problematic than the first one, not only because the latent trait would have to be estimated, but also because it might be estimated using the aberrant responses that the procedure is designed to detect.

A third assumption used in the simulation was that those having prior access to an item could be thought of as a random sample from the general population of test takers. It may be more realistic to think that the test takers with an interest in memorizing the correct response to an item are those for whom the item would otherwise be a challenging one. In other words, those whose latent trait value is less than the item difficulty might be more likely to memorize the correct response to an item than those with higher latent trait values.

References

Haberman, S. J. (2008). *Outliers in assessments* (ETS Research Rep. No. RR-08-41). Princeton, NJ: ETS.

Montgomery, D. C. (2008). *Introduction to statistical quality control* (6th ed). New York: Wiley.

Shu.Z., Luecht,L., & Henson, R. (2011). Using the Deterministic, Gated Item Response Model Detecting Test Cheating. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans.