# *Journal of Computerized Adaptive Testing*

***Volume 1 Number 1***

***December 2012***


# Termination Criteria in Computerized Adaptive Tests: Do Variable-Length CATs Provide Efficient and Effective Measurement?


**Ben Babcock and David J. Weiss**



**DOI 10.7333/1212-0101001**

**The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing**

***www.iacat.org/jcat***

**ISSN: 2165-6592**




---

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

*Journal of Computerized Adaptive Testing*
*Volume 1, Number 1, December 2012*
DOI 10.7333/1212-0101001
ISSN 2165-6592

# Termination Criteria in Computerized Adaptive Tests: Do Variable-Length CATs Provide Efficient and Effective Measurement?

**Ben Babcock,** ***The American Registry of Radiologic Technologists***
**David J. Weiss,** ***University of Minnesota***

This simulation study examined a number of computerized adaptive testing (CAT) termination rules based on the item response theory framework. Results showed that longer CATs yielded more accurate trait estimation, but there were diminishing returns with a very large number of items. Standard error termination performed quite well in terms of both administering a small number of items and having high accuracy of trait estimation if the standard error level used was low enough, but it was sensitive to the item bank information structure. Change in estimated $\theta$ performed comparably to standard error termination, but was less sensitive to the bank information structure. Fixed-length CATs performed either slightly worse or comparable to their variable-length termination counterparts; previous findings stating that variable-length CATs are biased were the result of artifacts, which are discussed. Recommendations for CAT termination are provided.

Keywords: *CAT, adaptive testing, termination, item response theory, 3PL model, monte-carlo simulation*

Computerized adaptive tests (CATs) are becoming increasingly popular in numerous domains (Fliege et al., 2005; Simms & Clark, 2005; Triantafillou, Georgiadou, & Economides, 2007). Computer availability and advances in item response theory (IRT; Weiss & Yoes, 1991; De Ayala, 2009) make adaptive testing more feasible in applied settings. CATs tailor the test to each examinee in order to obtain accurate measurement across the entire latent trait continuum. CATs are also advantageous over non-adaptive tests because CATs can administer fewer items to examinees while maintaining the same quality of measurement as non-adaptive tests (Gibbons, et al., 2008, 2012; Weiss, 1982, 2011).

CATs require six main components: (1) a response model, (2) an item bank of pre-tested items, (3) an entry rule, (4) an item selection rule, (5) a scoring mechanism, and (6) a termination rule (Weiss & Kingsbury, 1984). IRT is usually the statistical framework for CAT because IRT has a variety of options for fulfilling these requirements. There has been substantial research on

IRT parameter estimation (e.g., Harwell, Stone, Hsu, & Kirisci, 1996), some research on CAT entry rules (e.g., Gialluca & Weiss, 1979), item selection (e.g., Hau & Chang, 2001), and scoring methods (e.g., Wang & Vispoel, 1998). Although studies have examined a few termination criteria (e.g., Dodd, Koch, & De Ayala, 1993; Gialluca & Weiss, 1979; Wang & Wang, 2001), no single study has thoroughly compared a large number of termination rules using multiple item banks. This study examined numerous termination rules with four item banks to determine which termination rules led to the best CAT latent trait estimation.

The relative lack of termination studies is problematic because termination rules are important to delivering good CATs. How a CAT terminates is the driving factor behind the number of items a CAT uses. The measurement efficiency goal of CAT necessitates measuring examinees well with a small number of items (Weiss, 2011). If a CAT has a termination rule that either terminates the CAT before there is good measurement or administers a large number of items, the CAT user has failed at achieving efficient and/or effective measurement.

There are two classes of CAT termination criteria: fixed-length and variable-length. Fixed-length termination rules end a CAT after giving some constant number of items. Variable-length CATs, however, have differing numbers of items depending on individual item response patterns. There are two main advantages of using variable-length termination: efficiency and quality of measurement. Efficiency in measurement is measuring an individual using relatively few items; this reduces the amount of time required to measure people and optimizes the use of an item bank (Weiss & Kingsbury, 1984). The quality of measurement goal is to measure people with high precision. Variable CAT termination rules administer more items to ensure that examinees are measured to desired degrees of precision (Weiss, 1982, 2011).

Depending on the rule used, variable-length CATs should provide equal or superior measurement quality to fixed-length CATs, in theory. Once a test score has a certain amount of measurement precision, adding a few more items does not substantially change the score's precision. This relationship between test length and quality can easily be seen by exploring how increasing test length affects reliability using the Spearman-Brown prophecy formula (Brown, 1910; Spearman, 1910). If a CAT achieves good measurement precision based on some criterion and terminates earlier than a fixed-length CAT, there should not be a large difference in measurement quality. A fixed-length CAT that is much longer (say, double or triple the number of items) than a variable-length CAT could lead to somewhat better measurement (Brown, 1910; Spearman, 1910). The longer CAT would, however, fail on the goal of efficiency in measurement. Variable-length CATs continue administering items only if a CAT has not yet achieved some quality metric.

There could be situations, however, where variable-length termination is disadvantageous compared to fixed-length termination. If a variable termination criterion terminates a CAT with very few items, measurement quality will likely be poor. It is also possible that an item bank does not have enough items to satisfy a stringent termination criterion. A variable-length CAT might administer too many items, whereas a fixed-length CAT would terminate as expected. It is also easier to maintain and implement quality control with fixed-length CAT than variable-length CAT, but at the cost of potential decreases in efficiency. Candidates may also view fixed-length CAT as being fairer. An organization could, however, manage candidate perceptions by communications to them that define "fairness" as equal measurement precision.

## CAT Termination Methods

There are numerous variable-length termination criteria for use in CAT. The most popular variable-length termination rule is standard error (SE) termination (Weiss & Kingsbury, 1984).

SE termination administers items until a trait estimate reaches a specified precision level, resulting in equiprecise examinee measurement. This method depends on having enough informative items to meet the required standard error and on there being relatively little person misfit. The empirical SE of a $\theta$ estimate in IRT when using maximum likelihood scoring is

$$\mathrm{SE}(\hat{\theta}_p) = \frac{1}{\sqrt{-\partial^2 \log L / \partial \theta^2}}, \tag{1}$$

where $\hat{\theta}_p$ is the examinee's current $\theta$ estimate and log $L$ is the log likelihood function (Samejima, 1977). The log likelihood function is defined as

$$\log L(x_{1p}, x_{2p}, ..., x_{np}) = \sum_{i=1}^{n} x_{ip} \log[P_i(\theta)] + (1 - x_{ip}) \log[Q_i(\theta)], \tag{2}$$

where $i$ is an item index, $n$ is the number of items to which a person has responded, $x_{ip}$ is the scored (0,1) response of examinee $p$ to item $i$, $P_i$ is the IRT probability of a person responding in the keyed direction, and $Q$ is $1 - P_i$ (Embretson & Reise, 2000, Ch. 7). When an examinee has a response pattern with high psychometric information (i.e., has answered questions with difficulties near true $\theta$), the log likelihood function will be steeply curved at the $\theta$ estimate. This curvature is reflected in a decrease in the SE at the $\theta$ estimate. Researchers have found that SE termination performs well for equiprecise estimation of $\theta$ in CAT if the information in an item bank allows an SE value to be attained at all levels of $\theta$ (Dodd, Koch, & De Ayala, 1989, 1993 Revuelta & Ponsoda, 1998; Wang & Wang, 2001).

A second variable-length termination rule that has been studied in a limited fashion is the minimum information (MI) termination rule. This rule states that a CAT should end when there are no items remaining in a test bank that can provide more than some minimal amount of psychometric information at the current $\theta$ estimate (Gialluca & Weiss, 1979; Maurelli & Weiss, 1981). Model-predicted Fisher item information $I_i$ is calculated by

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta) Q_i(\theta)}, \tag{3}$$

where $P'$ is the first derivative of the item response function (Samejima, 1977). The total bank information is simply the sum of all of the individual item information values conditional on $\theta$. Individuals whose responses contain high Fisher information will be measured well with a low SE. Most item selection criteria in CAT involve choosing items with high information at or near the current $\hat{\theta}$. The theory behind MI termination is that the test should terminate for efficiency's sake if there are no unadministered high-information items left in the bank.

Choi, Grady, and Dodd (2011) proposed a similar CAT stopping rule that they called the predicted standard error reduction criterion. This rule determines if the standard error would be reduced by administering additional items. Because psychometric information and the standard error are directly related (Samejima, 1977), this new CAT rule is related to the earlier-proposed minimum information criterion. The main difference is that the predicted standard error reduction criterion takes into account the amount of psychometric information in the items that have already been administered, while the minimum information criterion does not. Because of the conceptual similarity of the two methods, this study did not investigate the predicted standard error reduction criterion.

Some research has demonstrated that MI is an efficient termination rule (Brown & Weiss, 1977). Research by Dodd, Koch, and De Ayala (1989, 1993), however, found that MI performed slightly worse than a fixed SE rule when there was little or no psychometric information in the item bank at one of the extremes of the $\theta$ continuum. The values of MI that they used, however, were quite high (.45 to .5, versus .01 and .05 used by Brown & Weiss). The Dodd et al. studies also used extremely small item banks (30 items) that were polytomously scored. It is possible that the Dodd et al. results were highly dependent on the context of their study, so the present study included MI termination.

A third variable-length termination rule, which has received little research attention, is the change in $\theta$, or $\theta$ convergence, criterion. Taking more items in a CAT yields additional psychometric information, so a person's $\hat{\theta}$ changes after each item in a CAT. Changes in $\hat{\theta}$ are large at the beginning of a CAT but generally become smaller as the CAT converges (Weiss & Kingsbury, 1984; Weiss, 2011). The convergence of a $\theta$ estimate could provide a good stopping rule for a CAT. Hart, Cook, Mioduski, Teal, and Crane (2006) and Hart, Mioduski, and Stratford (2005) investigated a hybrid CAT termination rule that combined SE and $\theta$ convergence. The researchers concluded that $\theta$ convergence yielded good $\theta$ estimates for a CAT.

Efficiency and quality of measurement make variable-length termination rules attractive from a measurement perspective. The simplicity of fixed-length termination and its similarity to paper-and-pencil tests, however, have made it the most popular applied termination rule (Weiss, 1982; Gushta, 2003). Some researchers have even argued that variable-length CATs lead to poorer measurement than fixed-length CATs. Chang and Ansley (2003), citing Stocking (1987), suggested that variable-length CATs are biased. Stocking compared a 20-item fixed-length stopping rule with a model-estimated "true score" SE termination criterion. The research used a test bank of 120 items with peaked information. Results showed that fixed-length CATs generally had lower root mean square errors than the variable-length CATs. One issue with this study is that the results were not in the $\theta$ metric but in a metric using model-predicted "true score." Standard errors of people whose scores are near the middle of the score distribution are greater than those whose scores are at the ends of the score distribution in the "true score" metric. This is counter-intuitive in IRT because peaked exams contain the most psychometric information near the middle of the score distribution, which leads to lower standard errors (Weiss, 1982). While Stocking's study was well-designed, the results do not necessarily directly apply to a context where the $\theta$ scale is being used.

Yi, Wang, and Ban (2001) also claimed that variable-length termination rules lead to poorer person estimates than fixed-length termination rules. These authors compared a 30-item fixed-length CAT termination rule to fixed standard error termination. The results showed that the standard error termination rule yielded $\theta$ estimates that were conditionally biased toward the mean. A problem with these results is that they were based on Bayesian scoring using a normal (0, 1) prior to estimate $\theta$. Bayesian person scoring algorithms bias estimates toward the mean of the prior (e.g., Guyer, 2008; Stocking, 1987; Wang & Vispoel, 1998; Weiss & McBride, 1984), with the amount of bias a function of the amount of information provided by the item responses. The variable-length CATs used less than half the number of items as the fixed-length conditions, making the priors more influential on the variable-length CATs. The combination of Bayesian scoring and test length created a statistical artifact that favored fixed-length termination[1].

[1] This should not imply that Bayesian scoring rules are incompatible with variable-length CAT. CAT users should simply keep in mind that Bayesian scoring using traditional prior distributions (i.e., normal) generally regress $\theta$ estimates toward the mean of the prior. The present study did not use Bayesian scoring in order to avoid this well-known statistical occurrence from confounding comparisons between methods.

## Purpose

The present study examined whether fixed-length CATs measure better than variable-length CATs under conditions designed to control for alternative explanations. In addition to using a wider variety of termination criteria than previous studies, this study also used several item banks. Nearly every CAT study uses one, or possibly two, item banks that reflect the needs of a very specific testing program. The few studies that have used a variety of banks (e.g., Dodd et al., 1989, 1993) did not examine a wide variety of termination criteria. The composition of the item bank, particularly in terms of psychometric information, has an effect on CAT termination and the quality of CAT measurement. This study sought to paint a more complete picture by using both a large number of termination criteria and several different item banks with differing numbers of items and information shapes. This research did not, however, examine termination criteria in the context of classification testing; it assumed that the goal of CAT was to measure well across the entire $\theta$ continuum. For a review of methods for CAT termination in a classification context, see Thompson (2007).

This study used monte-carlo simulation to investigate various CAT termination rules. Several conditions of four basic termination rules (SE, MI, change in $\theta$, and fixed-length) and two combinations of SE and MI termination were examined. The combination conditions were designed to take advantage of precise measurement where it was possible and measurement efficiency when high precision was not possible. The fixed-length termination conditions administered the mean number of items from selected variable-length conditions in order to compare fixed-length CATs to variable-length CATs. Having a comparable number of items is important, as past studies have generally used many more items for fixed-length CATs. The four item banks in this study were used to examine the conditions under which a given termination rule was superior or inferior to other rules in its ability to recover true $\theta$. Content balancing and item exposure were not considered in this study in order to avoid confounding the results with these operational CAT constraints.

# Method

## Model

Although there are a wide variety of IRT models (DeAyala, 2009; Embretson & Reise, 2000), this study used the unidimensional 3-parameter logistic (3PL) IRT model for dichotomously scored items. The mathematical form of this model is

$$P(x_{ip}=1 \mid \theta_p, a_i, b_i, c_i) = c_i + (1-c_i)\frac{\exp[Da_i(\theta_p - b_i)]}{1+\exp[Da_i(\theta_p - b_i)]}, \quad (4)$$

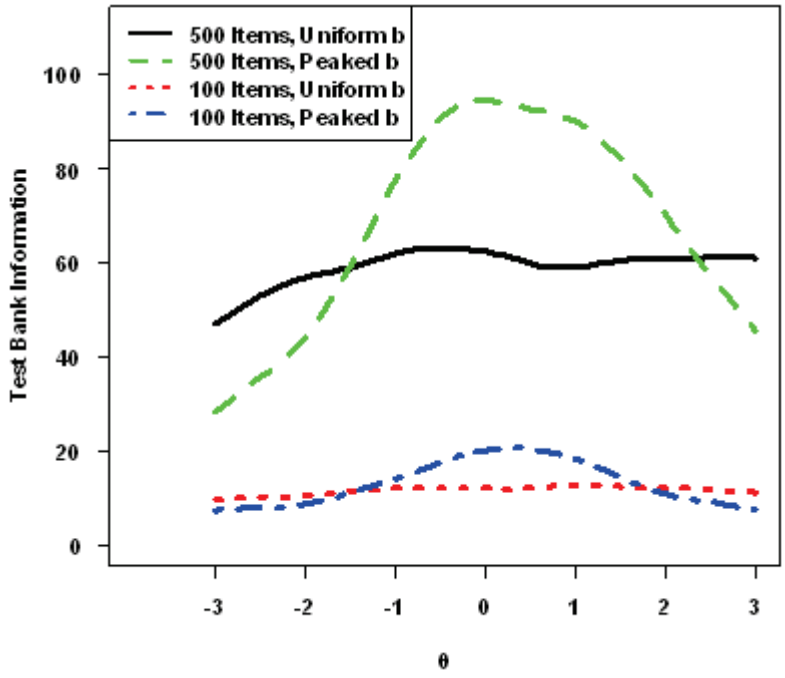
where $i$ is an item index, $p$ is a person index, $x_{ip}$ is a person's response to an item (1 for a keyed response, 0 for a non-keyed response), $D$ is the multiplicative constant 1.702, $a_i$ is the item discrimination parameter, $b_i$ is the item difficulty parameter, and $c_i$ is the lower asymptote (pseudo-guessing) parameter (Birnbaum, 1968; Weiss & Yoes, 1991; De Ayala, 2009).

## Item Banks

There were four item banks in this study: (1) a 500-item bank with uniform *b*s, (2) a 500-item bank with a peaked *b* distribution, (3) a 100-item bank with uniform *b*s, and (4) a 100-item bank with a peaked *b* distribution. Figure 1 contains the bank information functions for these four banks. These item banks had several notable features. First, information on the low end of $\theta$ was somewhat lower than at the high end of $\theta$. This occurred because the *c* parameter in the 3PL de-

creases the slope of an item response function in the low ranges of $\theta$; decreases in slope lead to lower item information. Second, the 500-item banks had a relatively high level of total information across the entire range of $\theta$. These banks could satisfy all of the SE criteria used as stopping rules for most of the $\theta$ range. An information value of 20.67, for example, corresponds to a model-predicted SE of 0.22 $[1/(20.67)^{1/2} = 0.22]$. Both of the 500-item banks had model-predicted information values greater than 21 for the entire $\theta$ continuum. Finally, the 100-item banks had much lower levels of total information. These item banks could not satisfy some of the lower SE termination criteria.

**Figure 1. Bank Information Functions for the Four Item Banks**




The *a* parameters for all four banks were generated from a log normal distribution with a log mean of –0.1 and a log SD of 0.3; this translated into a mean $a_i$ of 0.94 with a SD of 0.28. The *b* parameters for the two flat banks were distributed Unif(–4 , 4). The *b* parameters for the peaked banks came from a mixed distribution; 150 and 50 items in the large and small banks, respectively, came from a Unif(–4, 4) distribution. The remainder of the *b*s came from a N(0, 1) distribution. This mix of distributions gave the item banks a shape that mimicked item banks seen in practice, where most items were in the middle of the $\theta$ distribution with a slightly larger number of extreme items than expected from only a normal distribution. The *c* parameters for all item banks came from a mixture of uniform distributions, one Unif(0.100, 0.225) and the other Unif(0.075, 0.350). This mixture of distributions mimicked *c* parameters seen from real items in previous research (Chen & Ankenman, 2004; Wang, Hanson, & Lau, 1999).

## Data Generation and CAT Simulation

In order to evaluate the performance of the stopping rules on the $\theta$ continuum, this study used 13,000 simulees—1,000 at each of 13 evenly spaced points on $\theta$ from –3 to 3. Each $\theta$ had a model-predicted probability (Equation 4) of responding in the keyed direction to a given item. The response simulation compared this probability to a random number distributed Unif(0, 1). If

the probability of a keyed response was greater than the random number, the response was in the keyed direction (1). If the probability of a keyed response was less than the random number, the response was in the non-keyed direction (0). Each item bank had its own simulated dataset. All termination criteria, however, used the same simulated data within an item bank. POSTSIM3, a post-hoc CAT simulation program (POSTSIM3, 2008), simulated the CATs after the generation of the data sets. POSTSIM3 allows users to conduct a simulated CAT based on specified item parameters and a person's full set of item responses. Estimated CAT $\theta$ values for each termination condition from POSTSIM3 were then compared to the true $\theta$s.

## CAT Conditions

1. *Starting rule*. All examinees started with an initial $\theta$ of 0. Using the same starting point for every CAT eliminated chance starting variation from affecting the results of the study.
2. *Item selection rule*. This study used maximum Fisher information to select items for the CAT. POSTSIM3 selected the unadministered item with the highest information at the current $\theta$ estimate for each simulee. This rule maximizes the efficiency of the CATs by reducing the SE of $\theta$ (Equation 1) as quickly as possible.
3. *$\theta$ estimation.* All conditions used maximum likelihood (ML) estimation for $\theta$. If a simulee does not have a mixed response vector (i.e., has all keyed or all non-keyed responses), ML scoring does not yield a finite $\theta$ estimate. The CAT algorithm increased (for all keyed responses) or decreased (for all non-keyed responses) the item selection location by a fixed step size of 0.5. This rule obtained a mixed response vector relatively quickly, and ML scoring then began. This research did not use Bayesian scoring methods because previous research has demonstrated that Bayesian $\theta$ estimation methods produce biased $\theta$ estimates when true $\theta$ is extreme and for short CATs (Guyer, 2008; Stocking, 1987; Wang & Vispoel, 1998; Weiss & McBride, 1984).
4. *Termination rules*. For each item bank's data set, the same simulated response data were run 14 times, with each run using a different termination rule. The CAT terminated in each condition when the following condition was met:
   (1) SE($\theta$) was below 0.385 (analogous to a reliability[2] of 0.85) with a 100-item maximum, labeled as "SE1" in the tables.
   (2) SE($\theta$) was below 0.315 (reliability of 0.90) with a 100-item maximum (SE2).
   (3) SE($\theta$) was below 0.220 (reliability of 0.95) with a 100-item maximum (SE3).
   (4) All non-administered items at the current $\hat{\theta}$ had less than 0.2 Fisher information with a 100-item maximum (MI1).
   (5) All non-administered items at the current $\hat{\theta}$ had less than 0.1 Fisher information with a 100-item maximum (MI2).
   (6) All non-administered items at the current $\hat{\theta}$ had less than 0.01 Fisher information with a 100-item maximum (MI3).
   (7) Either the SE($\theta$) was below 0.315 or all non-administered items had less than 0.1 information at $\hat{\theta}$, with a 100-item maximum, whichever occurred first (MI,SE1).

[2] In classical test theory, the standard error of measurement (SEM) is approximated with the equation SEM = $s_x$ $(1-\rho_{xx})^{1/2}$, where $s_x$ is the standard deviation of the observed scores and $\rho_{xx}$ is the reliability estimate. Assuming that the standard deviation of $\theta$ is 1, specifying an SEM of 0.385 results in a reliability of about 0.85 using this very rough approximation. See Daniel (1999) for a more thorough discussion concerning reliability and "local reliability," which is more analogous to IRT's SEM.

(8) Either the SE($\theta$) was below 0.220 or all non-administered items had less than 0.01 information at the current $\hat{\theta}$, with a 100-item maximum (MI,SE2).

(9) Absolute change in $\hat{\theta}$ was less than 0.05 with an 11-item minimum and a 100-item maximum. The minimum number of items was to ensure that the CAT did not terminate prematurely in the event that a simulee answered several consecutive items correctly. This amount of change was just under 1% of the ±3 $\theta$ scale in which examinees generally fall ($\Delta\theta$1).

(10) Absolute change in $\hat{\theta}$ was less than 0.02 with an 11-item minimum and a 100-item maximum. This step size was only 1/3 of 1% of the ±3 $\theta$ scale in which examinees generally fall ($\Delta\theta$2).

(11) Fixed-length CAT using the mean number of items from Condition 2 (SE2F).

(12) Fixed-length CAT using the mean number of items from Condition 5 (MI2F).

(13) Fixed-length CAT using the mean number of items from Condition 7 (MI,SE1F).

(14) Fixed-length CAT using the mean number of items from Condition 9 ($\Delta\theta$1F).

The fixed-length conditions were used to compare fixed-length CATs with variable-length CATs using a comparable number of items.

## Dependent Variables

1. *Length of the CAT*. This was simply the number of items the CAT required to terminate. This dependent variable is a measure of the efficiency of variable-length CATs.
2. *Bias*. This statistic is the mean signed difference between the CAT-estimated $\theta$ and true $\theta$. Bias was calculated for each of the 13 $\theta$ values by

$$\text{bias}(\theta) = \frac{\sum_{p=1}^{N_\theta} (\hat{\theta}_p - \theta_p)}{N_\theta} , \tag{5}$$

   where $N_\theta$ is the number of people at a $\theta$ point, $\theta_p$ is a person's true $\theta$, and $\hat{\theta}_p$ is a person's CAT-estimated $\theta$.
3. *Root mean squared error* (*RMSE*). This statistic is a measure of the absolute difference between the CAT-estimated and true $\theta$*s*,

$$\text{RMSE}(\theta) = \sqrt{\frac{\sum_{p=1}^{N_\theta} (\hat{\theta}_p - \theta_p)^2}{N_\theta}} . \tag{6}$$

The overall (i.e., not conditional) dependent variables were calculated using weighted results such that the results reflected $\theta$ following a normal distribution with a mean of 0 and a variance of 1. The unweighted results were very similar to the normal (0,1) results, so only the normally-weighted results appear in the tables.

# Results

## Bank 1: Uniform *b* Item Bank of 500 Items

Table 1 contains the major results from all item banks, combined across $\theta$ levels. Several of

the conditions produced identical results. Conditions MI1, MI2, MI3, and MI2F all gave the maximum or nearly the maximum number of items to examinees for Banks 1 and 2 because these large item banks had a large number of items that could satisfy the minimum information criterion. The mean bias for these conditions was always 0 and the mean RMSE was always 0.16. These results were trivial and, thus, were omitted from Table 1. Conditions SE2F and MI,SE1F were also equivalent in the first two banks because these CAT conditions were fixed length with the same number of items. Conditions SE2F and MI,SE1F and Conditions SE3 and MI,SE2 were identical because Conditions MI,SE1 and MI,SE2 always reached SE termination before reaching minimum information termination with the first two banks. Thus, Conditions MI1, MI2, MI3, MI,SE1, MI,SE2, MI2F, and MI,SE1F were eliminated from the analyses for Banks 1 and 2.

SE below 0.385 (SE1), SE below 0.315 (SE2), and $\Delta\theta$ less than 0.05 ($\Delta\theta$1) administered the fewest items among the variable-length conditions. The results conditional on $\theta$ (available in the supplementary data file) showed the middle ranges of $\theta$ using slightly fewer items than extreme $\theta$s for the variable-length CATs. This occurred because the starting value for each person was $\theta = 0$, which was close to the person's true $\theta$ in the middle ranges; consequently, the test terminated a few items more quickly. Overall, the spread of the number of items administered was low, with standard deviations near or below 1% of the item bank size.

The mean bias for all of the variable-length conditions was very close to 0. The slight positive mean bias in SE1 occurred because this termination method did not estimate low values of $\theta$ very well. Figure 2 shows the bias conditional on $\theta$ for SE1, SE2, $\Delta\theta$1, SE2F, and $\Delta\theta$1F. The variable-length Conditions SE2 and $\Delta\theta$1 had very little conditional bias across $\theta$. The fixed-length Conditions SE2F and $\Delta\theta$1F also had very little bias. The poor results for SE1 were due to CATs terminating too early in the low ranges of $\theta$.

The RMSEs yielded some interesting results concerning the length of a CAT and the accuracy of $\theta$ estimation. RMSE across all methods was strongly related to the number of items administered; Figure 3 illustrates this relationship. Mean RMSE decreased quickly between 0 and 35 items. The point on the far right shows that the RMSE decreased more slowly once the test was over 40 items long. As seen in Table 1, both Condition SE2 and Condition $\Delta\theta$1 had slightly lower RMSE than their fixed-length counterparts SE2F and $\Delta\theta$1F.

The conditional RMSE values in Figure 4a indicate that the SE2 termination criterion had lower RMSE in the low regions of $\theta$ than the fixed-length SE2F. The SE termination criterion, when set strictly enough, administered more items to individuals in the extremes of the $\theta$ continuum who were not measured very well with only a few items.

Conditions $\Delta\theta$1 and $\Delta\theta$2 performed relatively well in terms of test length and bias. Although the $\Delta\theta$1 condition did not measure quite as well as the SE2 condition at the low end of $\theta$ (see Figure 4a), this simple method yielded $\theta$s with good measurement properties, essentially equivalent to or better than SE2 throughout almost all of the $\theta$ range.

## Bank 2: Peaked *b* Item Bank of 500 Items

Table 1 also contains the results combined across $\theta$ levels from Bank 2, which had fewer items with high information at the extremes of $\theta$. Conditions SE1, SE2, and $\Delta\theta$1 once again administered the fewest items. All of the conditions were relatively unbiased when combined across $\theta$, except for SE1. This positive bias occurred because SE1 did not estimate low values of $\theta$ very well, a trend similar to the results from Bank 1.

**Table 1. Normal Distribution Weighted Test Length, Bias, and RMSE for All Item Banks, By Condition**

| Statistic | Condition | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SE1 | SE2 | SE2F | SE3 | MI1 | MI2 | MI2F | MI3 | MI, SE1 | MI, SE1F | MI, SE2 | $\Delta\theta$1 | $\Delta\theta$1F | $\Delta\theta$2 |
| **Item Bank 1: 500 Items, Uniform *b*** | | | | | | | | | | | | | | |
| Mean Length | 7.61 | 11.98 | 12.00 | 30.48 | 100.0 | 100.0 | 100.0 | 100.0 | 11.98 | 12.00 | 30.48 | 14.44 | 14.00 | 26.67 |
| SD Length | 2.08 | 2.70 | --- | 4.54 | --- | --- | --- | --- | --- | --- | --- | 2.88 | --- | 6.12 |
| Mean Bias | 0.04 | 0.02 | 0.01 | 0.01 | --- | --- | --- | --- | --- | --- | --- | 0.01 | 0.01 | 0.01 |
| Mean RMSE | 0.42 | 0.33 | 0.37 | 0.22 | --- | --- | --- | --- | --- | --- | --- | 0.30 | 0.32 | 0.24 |
| **Item Bank 2: 500 Items, Peaked *b*** | | | | | | | | | | | | | | |
| Mean Length | 8.07 | 12.35 | 12.00 | 29.74 | 100.0 | 100.0 | 100.0 | 100.0 | 12.35 | 12.00 | 29.74 | 14.79 | 15.00 | 29.46 |
| SD Length | 2.14 | 2.89 | --- | 5.69 | --- | --- | --- | --- | --- | --- | --- | 3.09 | --- | 6.23 |
| Mean Bias | 0.06 | 0.03 | 0.03 | 0.01 | --- | --- | --- | --- | --- | --- | --- | 0.02 | 0.02 | 0.01 |
| Mean RMSE | 0.44 | 0.34 | 0.37 | 0.22 | --- | --- | --- | --- | --- | --- | --- | 0.32 | 0.32 | 0.23 |
| **Item Bank 3: 100 Items, Uniform *b*** | | | | | | | | | | | | | | |
| Mean Length | 16.30 | 46.74 | 47.00 | 100.0 | 23.49 | 35.30 | 35.00 | 65.61 | 30.45 | 30.00 | 65.61 | 15.35 | 15.00 | 23.66 |
| SD Length | 3.55 | 26.77 | --- | --- | 2.92 | 3.94 | --- | 5.41 | 4.29 | --- | 5.41 | 2.86 | --- | 3.67 |
| Mean Bias | 0.02 | 0.01 | 0.00 | --- | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 0.01 |
| Mean RMSE | 0.42 | 0.33 | 0.31 | --- | 0.36 | 0.33 | 0.33 | 0.31 | 0.34 | 0.34 | 0.31 | 0.42 | 0.43 | 0.36 |
| **Item Bank 4: 100 Items, Peaked *b*** | | | | | | | | | | | | | | |
| Mean Length | 13.50 | 28.40 | 28.00 | 100.0 | 35.12 | 53.22 | 53.00 | 79.41 | 23.03 | 23.00 | 77.04 | 14.89 | 15.00 | 26.68 |
| SD Length | 6.13 | 20.51 | --- | --- | 8.87 | 11.81 | --- | 8.05 | 5.93 | --- | 9.25 | 2.97 | --- | 6.05 |
| Mean Bias | 0.03 | 0.01 | 0.01 | --- | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | 0.02 | 0.00 | 0.03 | 0.03 | 0.01 |
| Mean RMSE | 0.40 | 0.32 | 0.31 | --- | 0.30 | 0.27 | 0.26 | 0.25 | 0.33 | 0.33 | 0.25 | 0.40 | 0.40 | 0.32 |

**Figure 2. Conditional Bias for Selected Conditions in Item Bank 1**

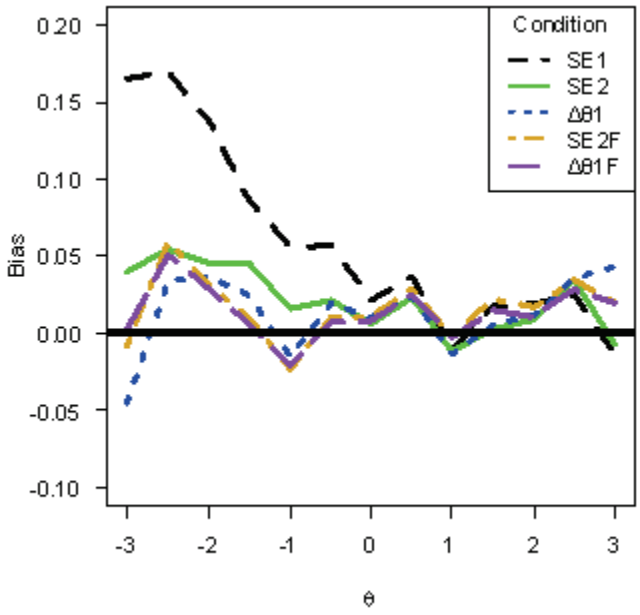



**Figure 3. Mean RMSE By Mean Test Length for Item Bank 1**

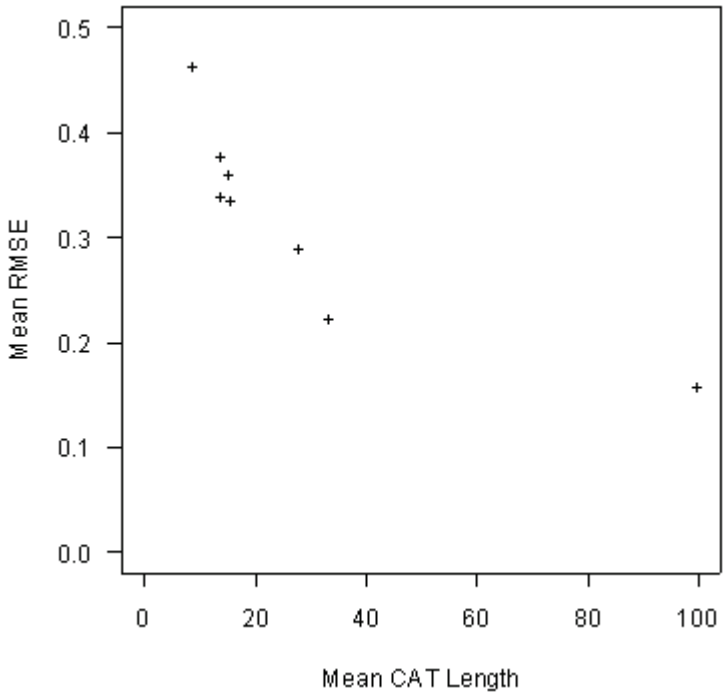
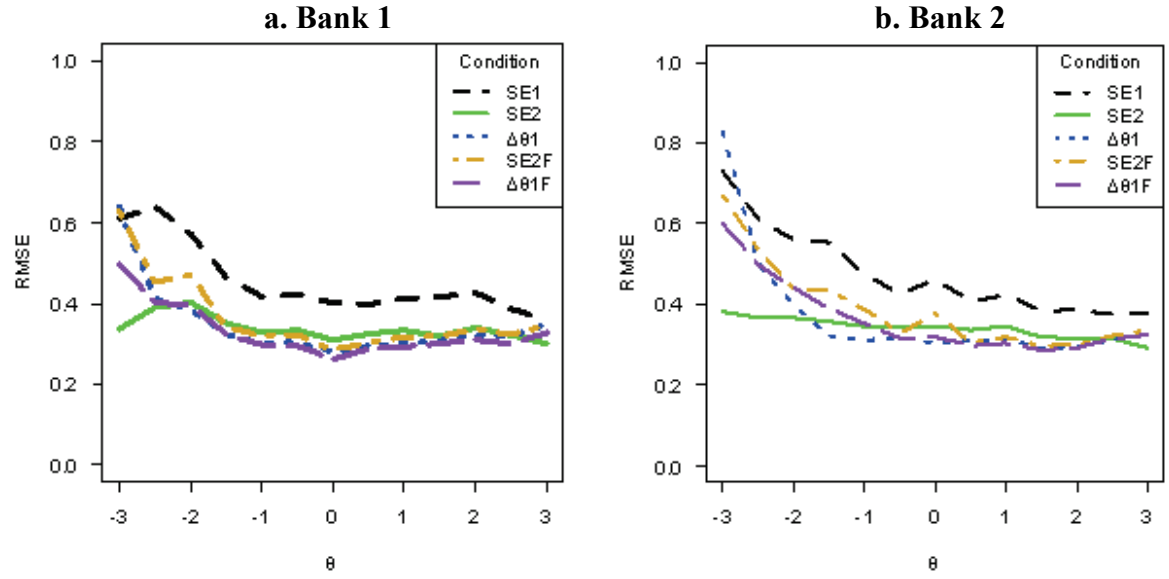


**Figure 4. Conditional RMSE for Selected Conditions in Item Banks 1 and 2**




The RMSEs from Bank 2 yielded results that were similar to those of Bank 1. The RMSE was strongly related to the number of items administered. The conditions that administered the largest number of items all had low RMSE across $\theta$. The conditions that administered fewer items had larger RMSE values for low $\theta$s. Figure 4b displays the conditional RMSE values for Conditions SE1, SE2, $\Delta\theta 1$, SE2F, and $\Delta\theta 1$F for Item Bank 2. Condition SE2 had a much lower RMSE for $\theta < -1.0$ than its fixed-length counterpart, SE2F. The SE termination criterion administered more items to people in the low ranges of $\theta$, where more items were needed to measure individuals well, so the variable-length SE2 condition performed better than the fixed-length condition SE2F. $\Delta\theta 1$, however, had virtually identical RMSE compared to its fixed-length counterpart, $\Delta\theta 1$F. The change in $\theta$ conditions ($\Delta\theta 1$ and $\Delta\theta 2$) performed similarly to CATs of comparable length that were terminated by SE or fixed length.

## Bank 3: Uniform *b* Item Bank of 100 Items

Table 1 also contains the weighted mean results for Bank 3. The SE termination criteria administered more items in this item bank because there was not a large number of highly discriminating items at every point on $\theta$. The SE3 condition administered the entire bank because this bank did not have sufficient Fisher information to support a termination criterion of SE = 0.22; results for this termination option are not shown in Table 1 for Bank 3. The minimum information criteria conditions administered fewer items than in Banks 1 and 2 for the same reason. The number of items administered for the $\Delta\theta$ conditions were not greatly affected by the change in the item bank. The $\Delta\theta$ conditions administered the fewest items overall for this item bank.

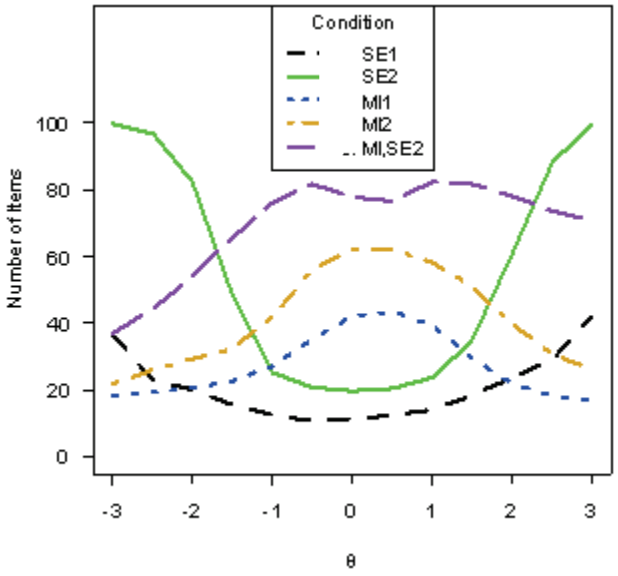
All of the conditions yielded $\theta$ estimates with biases near zero. All conditions had conditional RMSEs that were relatively close to the mean RMSE values. Conditions $\Delta\theta 1$, $\Delta\theta 2$ and $\Delta\theta 1$F had slightly higher conditional RMSE at $\theta = -3$. Giving a very large number of items from this small item bank did not result in large decreases in RMSE in the same way as did giving more items from the large bank. The large number of additional items that the CATs administered did not provide much information about the examinees.

Variable-length CATs in this item bank performed quite similarly to their fixed-length counterparts. The mean bias and RMSE were virtually identical between the variable-length CATs and the comparable fixed-length CATs.

## Bank 4: Peaked *b* Item Bank of 100 Items

For Bank 4, the SE3 condition again administered the entire bank for nearly every examinee, so these results are not shown in Table 1. The conditions that administered the fewest items for Bank 3 also administered the fewest items for Bank 4. Figure 5 is a plot of the conditional mean number of items administered across $\theta$ for selected conditions; conditions with substantial overlap or low conditional variability are not included. The number of items administered varied widely within some conditions. These differences occurred because of the greater information available in the middle of the $\theta$ distribution for this smaller item bank. Variable-length conditions requiring a SE cutoff (SE1 and SE2) administered many more items at the extremes of $\theta$ because the item bank often did not have enough items to fulfill the termination criterion. These differences were much greater than in the larger item banks. The MI criterion conditions (MI1 and MI2) administered fewer items in the extremes of $\theta$ because of a lack of informative items in this region.

**Figure 5. Conditional Mean Number of Items Administered for Selected Conditions for Item Bank 4**




The trends for conditional bias were quite similar across all conditions, varying around zero. The conditional RMSEs were slightly lower in the middle of $\theta$, and the conditions that administered more items had slightly lower RMSE overall. The fixed- and variable-length CATs of comparable length performed similarly, with the fixed-length conditions slightly out-performing their variable-length counterparts.

## Discussion and Conclusions

Several trends emerged when examining the results across item banks. As expected, CATs that administered more items yielded better $\theta$ estimates. The gains in accuracy were largest when adding items to short tests. Adding items to already long tests (e.g., 50 items) did not, however, yield sizable gains in $\theta$ estimation accuracy. Second, CATs that were too short (e.g., mean length less than 10 items) did not give good $\theta$ estimates in low ranges of $\theta$; the large conditional bias and RMSE values for the SE < 0.385 termination CATs demonstrated this. A CAT using dichotomously scored items should administer some minimum number of items, such as 10 to 15, before terminating, based on how very short CATs functioned in this study. CAT administrators using SE termination should use a standard error equal to or smaller than 0.315 for accurate $\theta$ estimation. Third, the variable termination criteria that performed the best when taking test length and accuracy into consideration were the conditions that used SE below 0.315 as part of the termination rule. These rules estimated low $\theta$ values more accurately than their fixed-length counterparts. Change in $\theta$, a relatively new termination rule, performed almost as well as the SE conditions and was less affected by changes in the item bank information structure. Finally, using minimum information termination alone administered too many items for large item banks.

One clear conclusion was that, contrary to claims in the literature (Chang & Ansley, 2003; Yi, Wang, & Ban, 2001), variable-length CATs generally performed equivalently to, or slightly better than, their fixed-length counterparts in terms of bias and RMSE. Previous results were a statistical artifact due to differences in the number of items administered and the $\theta$ estimation algorithm. First, fixed-length CATs in previous research were generally much longer than variable-length CATs. This means that fixed-length CATs in these studies utilized more Fisher information (Revuelta & Ponsoda, 1998), thus giving fixed-length tests an advantage in reducing the bias inherent in $\theta$ estimation for short tests. Second, many of these studies used Bayesian $\theta$ estimation. Numerous studies have shown that Bayesian scoring produces conditionally biased results, particularly when tests are short (Guyer, 2008; Stocking, 1987; Wang & Vispoel, 1998). Previous results claiming that variable-length CATs are biased were due to Bayesian estimation techniques combined with variable-length CATs that were substantially shorter than their fixed-length counterparts. The trend found in this study was that, no matter how a CAT is terminated, using too few items has a negative effect. This was especially true for the large item banks, where the SE = .385 condition terminated in a mean of less than 10 items. The $\theta$ estimates from this condition were highly biased in the low ranges of $\theta$ and had high RMSEs.

### Termination Recommendations

When using a large item bank, terminating with a strict (i.e., low) standard error yielded measurements of good quality. If a practitioner has an item bank that can support a relatively low standard error across the entire $\theta$ continuum or the vast majority of the $\theta$ continuum, the fixed standard error rule will result in equally good measurement for all examinees. This is an extremely desirable property in measurement. One context where equal quality in measurement is important is in admissions testing, where some institutions might be accepting more students scoring at the low end of the continuum, but other institutions might concentrate on the upper end of the continuum. Equal quality in measurement might be particularly useful from a practical standpoint when attempting to explain to candidates how tests of differing lengths are fair. If an organization can cast "fairness" as equal quality in measurement, then the angst of the examinees who listen to the message would likely be lower. If a psychometrician's item bank can support

an equal and low standard error of measurement rule, then the psychometrician should very strongly consider using standard error termination.

A useful solution in practice for variable CAT termination is to use one or more variable termination criteria in combination with a minimum number of items constraint. Based on this research, 10 to 15 items may be a *minimum* number of items for variable-length CAT termination with dichotomously scored items, depending on the precision needs of the test user and the discriminations of the items in the bank. Variable termination rules would supplement the minimum items rule by administering more items to people who are still not measured well. This would ensure stability of the $\theta$ estimates and fulfill the efficiency and precision goals of CAT users.

One effective combination that this study used is combining a standard error termination rule and a minimum information rule. The CAT would, thus, administer items until an individual was either measured well or there were no bank items left that could measure a person well. This method has advantages particularly for small (but also certain large) item banks with peaked information functions. The stopping rule combination stops quickly in areas of the item bank where people cannot be measured well (i.e., low total information areas) but continues to administer items in areas where the item bank can measure people well (high total information areas). While minimum information alone is good for very low total information banks and SE termination alone is good for high total information banks, combining the two rules yields a stopping rule suitable for virtually any item bank information structure.

The change in $\theta$ criterion, a relatively new termination criterion, performed comparably to other termination criteria with similar mean numbers of items. This method might also be a viable supplement to standard error termination when an item bank does not permit a given standard error to be reached in certain ranges of $\theta$, which can occur in peaked-information item banks. This method, furthermore, was not affected nearly as much by the information structure of the item banks as were the other methods. The standard deviation in number of items administered was reasonable for even the small item banks, where other termination rules varied quite widely. Change in $\theta$ may be a good option that practitioners could apply across a wide variety of item banks with different information structures. This could be particularly good for certification organizations, which often have a few high-volume entry exams with large item banks along with a variety of advanced exams with low examinee volume and smaller item banks. The usefulness of the change in $\theta$ criterion supports results obtained by Hart et al. (2005, 2006) using polytomous items.

This research does have limitations. The most notable limitation is that this study did not control for item exposure or content balancing. It is probable that techniques controlling for item exposure and/or content balancing—both issues in some CAT applications—would increase the minimum number of items required for a CAT to give a desired level of measurement precision.

CAT is a good way to measure accurately and simultaneously increase test efficiency. This research demonstrated that a wide variety of termination criteria work well if a minimally sufficient number of items is used. CATs can dramatically reduce the number of items required for accurate measurement over non-adaptive methods, and additional research will further delineate factors affecting the effectiveness CATs.

## References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries* (Research Rep. No. 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program. Available from www.iacat.org/biblio

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296–322.

Chang, S.-W., & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40,* 71–103. *CrossRef*

Chen, Y.-Y., & Ankenman, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement, 41,* 149–174. *CrossRef*

Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement, 71,* 37–53. *CrossRef*

Daniel, M. H. (1999). Behind the scenes: Using new measurement methods on DAS and KAIT. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement* (pp. 37–63). Mahwah, NJ: Lawrence Erlbaum Associates.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: Guilford Press.

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13,* 129–143. *CrossRef*

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement, 53,* 61–77. *CrossRef*

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research, 14*, 2277–2291. *CrossRef*

Gialluca, K. A., & Weiss, D. J. (1979). *Efficiency of an adaptive inter-subset branching strategy in the measurement of classroom achievement* (Research Report 79-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program. Available from www.iacat.org/biblio

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*, 49–58. *CrossRef*

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). The CAT-DI: A computerized adaptive test for depression. *Archives of General Psychiatry, 69 (11)*,1104-1112. *CrossRef*

Gushta, M. M. (2003, May). *Standard-setting issues in computerized-adaptive testing.* Paper presented at the Annual Conference of the Canadian Society for Studies in Education, Halifax, Nova Scotia.

Guyer, R. D. (2008). *Effect of early misfit in computerized adaptive testing on the recovery of theta.* (Doctoral Dissertation, University of Minnesota). Available from www.iacat.org/biblio

Hart, D. L., Cook, K. F., Mioduski, J. E., Teal, C. R., & Crane, P. K. (2006). Simulated computerized adaptive test for patients with shoulder impairments was efficient and produced valid measures of function. *Journal of Clinical Epidemiology, 59,* 290–298. *CrossRef*

Hart, D. L., Mioduski, J. E., & Stratford, P. W. (2005). Simulated computerized adaptive tests for measuring functional status were efficient with good discriminant validity in patients with hip, knee, or foot/ankle impairments. *Journal of Clinical Epidemiology, 58,* 629–638. *CrossRef*

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20,* 101–125. *CrossRef*

Hau, K.-T., & Chang, H.-H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38,* 249–266. *CrossRef*

Maurelli, V., & Weiss, D. J. (1981). *Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries* (Research Rep. No. 81-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory. Available from www.iacat.org/biblio

POSTSIM3: Post-hoc simulation of computerized adaptive testing (2008) [Computer software]. St. Paul, MN: Assessment Systems Corporation.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35,* 311–327. *CrossRef*

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1,* 233–247. *CrossRef*

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment, 17,* 28–43. *CrossRef*

Spearman, C. C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295.

Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. *Applied Psychology: An International Review, 36,* 263–277. *CrossRef*

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment, Research, & Evaluation, 12*. Available online: http://pareonline.net/getvn.asp?v=12&n=1

Triantafillou, E., Georgiadou, E., & Economides, A. A. (2008). The design and evaluation of a computerized adaptive test on mobile devices. *Computers & Education, 50, 4,* 1319-1330. *CrossRef*

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25,* 317–331. *CrossRef*

Wang, T., Hanson, B. A., & Lau, C.-M. A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement, 23,* 263–278. *CrossRef*

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35,* 109–135. CrossRef

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473–492. CrossRef

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences, 2,* 1–27. Available from https://journals.uair.arizona.edu/index.php/jmmss/article/view/12351

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21,* 361–375. CrossRef

Weiss, D. J., & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8,* 273–285. CrossRef

Weiss, D. J., & Yoes, M. E. (1991). Item response theory. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (pp. 69–95). Boston: Kluwer Academic.

Yi, Q., Wang, T., & Ban, J.-C. (2001). Effects of scale transformation and test-termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement, 38,* 267–292. CrossRef

## Supplementary Data

The Supplementary Data file for this article contains the following data:

- Item parameters for the four item banks.
- Bank information functions for the four banks (Figure 1).
- Numerical values for Figures 2–5.
- Conditional means and SDs of CAT length and bias, and RMSE and squared bias for selected termination conditions and all four banks.

This file can be requested from the Editor, djweiss@umn.edu.

## Acknowledgments

Parts of this paper were presented at the 2009 GMAC® Conference on Computerized Adaptive Testing. The first author was the sole correspondent and appeared as the sole author throughout the peer review process in order to ensure a truly blind review. The conclusions, discussions, and views contained in this article are not necessarily the official position of The American Registry of Radiologic Technologists.

## Author Address

Ben Babcock, The American Registry of Radiologic Technologists®, 1255 Northland Drive, St. Paul, MN 55120, U.S.A. Email ben.babcock@arrt.org.