Applied Psychological Measurement
2015, Vol. 39(4) 278–292
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621614561314
apm.sagepub.com
SAGE

Stochastic Curtailment in Adaptive Mastery Testing: Improving the Efficiency of Confidence Interval–Based Stopping Rules

Haskell Sie[1], Matthew D. Finkelman[2], Jay Bartroff[3], and Nathan A. Thompson[4]

Abstract

A well-known stopping rule in adaptive mastery testing is to terminate the assessment once the examinee's ability confidence interval lies entirely above or below the cut-off score. This article proposes new procedures that seek to improve such a variable-length stopping rule by coupling it with curtailment and stochastic curtailment. Under the new procedures, test termination can occur earlier if the probability is high enough that the current classification decision remains the same should the test continue. Computation of this probability utilizes normality of an asymptotically equivalent version of the maximum likelihood ability estimate. In two simulation sets, the new procedures showed a substantial reduction in average test length while maintaining similar classification accuracy to the original method.

Keywords

item response theory, adaptive mastery testing, stochastic curtailment, ability confidence interval

In the past several decades, computerized adaptive tests (CATs) have received much attention in educational and psychological research due to their efficiency in achieving the goal of assessment, whether it is to estimate the latent trait of test takers with high precision or to accurately classify them into one of several latent classes. In the latter case, the adaptive nature of CATs is used in educational testing to make inferences about the location of examinees' latent ability

[1]American Institutes for Research, Washington, DC, USA
[2]Tufts University, Boston, MA, USA
[3]University of Southern California, Los Angeles, USA
[4]Assessment Systems, Woodbury, MN, USA

Corresponding Author:
Haskell Sie, American Institutes for Research, 1000 Thomas Jefferson Street NW, Washington, DC 20007-3835, USA.
Email: hsie@air.org

relative to one or more prespecified cutoff points along the ability continuum. When there is only one cutoff point and two proficiency groups, this type of CAT is commonly referred to as adaptive mastery testing (AMT; Spray & Reckase, 1996; Weiss & Kingsbury, 1984). In this setting, an examinee will pass the test and be declared a ''master'' if evidence collected based on test responses indicates that the latent ability is above the threshold for mastery.

Among other advantages of AMT over conventional paper-and-pencil tests is the fact that different examinees might receive different sets of test items that best suit their ability levels. When test items are selected adaptively to match examinees' true ability, high-performing examinees will not be given easy items unnecessarily. Similarly, low-performing examinees will not be under undue pressure when trying to solve very difficult items. In any implementation of AMT, the choice of classification rule plays a vital role. Several methods that have been proposed include the sequential probability ratio test (SPRT; Spray, 1993; Thompson, 2011; Weissman, 2007), the sequential Bayes (SB) approach (Spray & Reckase, 1996), the generalized likelihood ratio (GLR) approach (Bartroff, Finkelman, & Lai, 2008; Thompson, 2011), and the confidence interval (CI) approach (Thompson, 2011; Weiss & Kingsbury, 1984). All of these approaches stem from the framework of sequential analysis (Wald, 1947), whereby inferences are continually made after each additional observation is obtained. Ideally, inferential procedures based on sequential analysis should be allowed to continue indefinitely until the desired confidence level is reached. In practice, however, there is usually a fixed upper bound n on the number of observations that the practitioners are willing to obtain. In the context of AMT, n is the maximum number of test items that any examinee can receive before a classification decision is made. When termination of the test is forced to occur after n observations are obtained, it is said that a *truncated sequential procedure* is being used.

Further improvements to sequential procedures for classification purposes have been proposed in the literature. Eisenberg and Ghosh (1980) discussed a modification rule called *curtailment* that allows for a sequential procedure to be stopped early if no subsequent observations can alter the final classification decision. By this specification, the curtailed version of a sequential procedure always yields the same classification results as the original procedure. Therefore, the two methods always have the same proportion of correct decisions (PCDs). However, the curtailed version has lower average test length (ATL) due to the possibility of early stopping. As a result, curtailed truncated sequential procedures could be useful in many different applications, including in clinical trials and educational testing. In educational testing, administering the fewest possible number of items while trying to make accurate classification decisions is important because a shorter test reduces the operational cost of assessment as well as alleviating the problem of overexposing items in high-stakes testing programs.

As a further improvement to the aforementioned idea of curtailment, a sequential procedure can also be halted during interim analyses when subsequent observations have only a small probability of altering the final results of analysis. Within this framework of *stochastic curtailment* that was first introduced by Lan, Simon, and Halperin (1982) in the context of clinical trials, several formulations are available depending on how the probability of whether the final results of analysis will change should the sequential procedure continue is calculated. Regardless of what approach is followed, many applications of stochastic curtailment have arisen in various fields. In the particular context of educational measurement, Finkelman (2008, 2010) used stochastic curtailment as a stopping rule for AMT in addition to the original truncated SPRT (TSPRT) and its curtailed version. It was shown that the stochastically curtailed TSPRT, albeit having slightly lower PCD than that of the original version, greatly improves the ATL whether or not the test is constrained from content balancing or exposure control perspectives.

Although curtailment and stochastic curtailment have both been applied in the AMT setting, their applications have only been studied when the TSPRT or the GLR approach is used as the stopping criterion. This fact limits their usefulness because the TSPRT or the GLR approach might not always be the stopping method of choice. Previous research has found that both the TSPRT and the GLR stopping criteria work best when test items are selected based on maximum Fisher information at the cut-off point (Reckase & Spray, 1994; Thompson, 2011). In terms of item bank structure, this requires that many items have high information at the cut-off score. If the item bank is relatively flat in terms of its information function, selecting test items based on maximum Fisher information at the interim ability estimate is more appropriate. In that case, the CI approach works best as a stopping rule due to the decreasing conditional standard error of measurement (CSEM; Thompson, 2007). In addition, because the TSPRT and the GLR stopping criteria are typically used in conjunction with the item selection method that maximizes Fisher information at the cut-off point, the same set of items will be selected for every examinee. Both stopping criteria are therefore prone to the problem of overexposure of test items. On the contrary, different sets of items will be selected for different examinees under the CI approach as it is usually coupled with an adaptive item selection method. Therefore, the distribution of item exposure rates might benefit from using a CI-based method, although a rather uneven distribution is still likely to be observed if item selection is based on Fisher information. Preference toward using the CI approach as the stopping rule over the TSPRT could also be attributed to the arbitrariness inherent in the latter. The TSPRT requires that the likelihood function be compared at 2 points close to the cut-off score, with the region along the latent trait continuum falling between those 2 points referred to as the ''indifference region'' (Thompson, 2007; Wald, 1947; Weissman, 2007). The TSPRT thus assumes that the practitioners are indifferent to the fact that the probability of misclassifying examinees is typically greater than the nominal Type I error rate within this region (Y.-C. Chang, 2005), an assumption that some practitioners might not be willing to make. It is also not clear how wide the indifference region should be, a factor that further affects how early the AMT can terminate. This dependence of the TSPRT on the width of the indifference region is explained by Thompson (2011).

The purpose of this article is to examine whether the CI stopping rule for AMT can benefit from the use of curtailment and stochastic curtailment. In particular, it is investigated whether curtailment and stochastic curtailment can substantially reduce the ATL of the CI stopping rule, without compromising classification accuracy. If so, the curtailed and stochastically curtailed versions of the CI stopping rule would be attractive options under the circumstances (outlined earlier) where the CI approach is preferable to the TSPRT and the GLR.

In constructing the CI for the person ability parameter, an interim ability estimate needs to be available. While the maximum likelihood (ML) estimator is commonly used for this purpose, it is sometimes not uniquely defined when the three-parameter logistic (3PL) model (explained below) is used, as it is in this article. Therefore, for the purposes of this study, the modified version of the ML estimator introduced by Chang and Ying (2009) will be used. This modified estimator is unique whenever it exists and is asymptotically equivalent to the original ML estimator. It is this modified ML estimator that will guide adaptation of curtailment and stochastic curtailment to the framework of the CI stopping rule.

Item Response Theory (IRT) and the Truncated CI Stopping Rule in AMT

IRT

In AMT, the classification decision for examinees is built upon the relationship between their latent ability and item responses. This relationship is commonly characterized via IRT models. One of the most common models used when test items are scored dichotomously (i.e., correct or incorrect) is the 3PL model. With U_i being a Bernoulli random variable that takes the value 1 if item i is answered correctly and 0 otherwise, the probability of an examinee with ability θ answering item i correctly is given by

$$P_i(\theta) \equiv P(U_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp\{-a_i(\theta - b_i)\}}, \tag{1}$$

where a_i is the item discrimination parameter, b_i is the item difficulty parameter, and c_i is the item guessing parameter. When all c_i in Model 1 are set to 0, the two-parameter logistic (2PL) model is obtained. In addition to having all zero guessing parameters, if all a_is in Model 1 have the same value, the one-parameter (1PL) model is obtained.

Using the model introduced earlier, the likelihood function upon observing an examinee's responses to the ith test item is given by $L(u_i|\theta) = \{P_i(\theta)\}^{u_i}\{Q_i(\theta)\}^{1-u_i}$, where $Q_i(\theta) = 1 - P_i(\theta)$. The likelihood function upon observing the responses to n test items is $L_n \equiv L(u_1, \ldots, u_n|\theta) = \prod_{i=1}^{n} L(u_i|\theta) = \prod_{i=1}^{n} \{P_i(\theta)\}^{u_i}\{Q_i(\theta)\}^{1-u_i}$, which is a consequence of the assumption that all test responses are independent given θ (i.e., the local independence assumption). The likelihood function can then be used to define the Fisher information for item i as $FI_i(\theta) = -E[\partial^2/\partial\theta^2\{\log L(U_i|\theta)\}] = Q_i(\theta)\{P_i(\theta) - c_i\}^2 \big/ P_i(\theta)(1 - c_i)^2 a_i^2$ as well as the Fisher information of a test of n items given by $FI^{(n)}(\theta) = -E[\partial^2/\partial\theta^2\{\log L_n\}] = \sum_{i=1}^{n} FI_i(\theta)$. It is well known that this Fisher test information is inversely proportional to the asymptotic variance of the ML ability estimate. Therefore, item selection algorithms in AMT and other CATs often attempt to maximize the Fisher test information.

Truncated CI Stopping Rule in AMT

The idea of using the CI stopping rule for classification of examinees in AMT originated from Weiss (1983). With $\hat{\theta}^{(k)}$ denoting the examinee's ML ability estimate computed using the first k test responses, a CI around $\hat{\theta}^{(k)}$ is constructed and its position relative to the test cutoff point θ_c is checked. If the lower end point of the CI is higher than θ_c, the test is terminated and the examinee is classified as a master. If the upper end point of the CI is below θ_c, the test is terminated and the examinee is classified as a nonmaster. If neither situation occurs, another test item is administered, the ability of the examinee is reestimated, and its associated CI is recomputed. As mentioned earlier, the asymptotic variance of the ML estimate of θ is inversely proportional to the test information. Therefore, the $100(1 - \alpha)\%$ CI for θ after the kth test item is administered can be approximated by

$$\left(\hat{\theta}^{(k)} - z_{1-\alpha/2}\left[FI^{(k)}(\theta)\right]^{-1/2}, \hat{\theta}^{(k)} + z_{1-\alpha/2}\left[FI^{(k)}(\theta)\right]^{-1/2}\right), \tag{2}$$

where $z_{1-\alpha/2}$ denotes the $1 - \alpha/2$ quantile of a standard normal random variable, and $FI^{(k)}(\hat{\theta}^{(k)})$ can be used in place of the unknown $FI^{(k)}(\theta)$.

The CI for θ can be constructed following Equation 2 even when k is small, as long as the examinee's response pattern already contains both correct and incorrect answers to guarantee the existence of a finite ML ability estimate. However, it is well understood that a CI formed using few observations is not highly reliable due to the large standard error of the estimate. With more observed test responses, the test information becomes larger because each term in the summation is nonnegative. Therefore, the CIs will shrink and give a more useful idea as to where the examinee's true ability actually lies. In practice, this is usually taken into account by imposing a minimum test length n_0, prior to which no classification decisions are made. Such a minimum test length could also be chosen according to the test blueprint. For example, on a mathematics test that consists of three sections (Algebra, Geometry, and Statistics), $n_0 = 15$ items could be imposed as a constraint if all three sections have to have a minimum of five items each. Once the minimum number of test responses has been observed, the CI for θ can be constructed and early stopping of the test might be invoked at any time before the maximum test length n is reached.

When the examinee's true ability is very close to the cutoff score, the desired early stopping scenario might not occur. The CI for θ constructed around $\hat{\theta}$ will potentially always contain θ_c, because $\hat{\theta}$ itself is likely to be close to the examinee's true ability, and therefore close to θ_c. In that case, the test will continue until all n items are administered. At that point, termination is inevitable, and the classification decision can be made simply by comparing the final ML ability estimate $\hat{\theta}^{(n)}$ with the cutoff score θ_c (Weiss & Kingsbury, 1984).

Curtailed and Stochastically Curtailed CI Stopping Rules in AMT

Curtailed CI Stopping Rule

As explained in the previous section, early stopping using the truncated CI stopping rule in AMT can occur any time at or after n_0, but before n items are administered if either the lower bound of the interim CI exceeds the cut-off score θ_c, or if the upper bound is less than θ_c. This sequential method can be stated mathematically as follows. Let $\mathrm{I}_k = (\underline{\theta}_k, \overline{\theta}_k)$ be the CI for θ at time k as defined in Equation 2. In addition, following Finkelman (2008, 2010), let D_T and K_T be the classification decision and stopping time, respectively, of the truncated CI stopping rule. Then, for any $k \in \{n_0, n_0 + 1, \ldots, n - 1\}$, set $K_T = k$ and $D_T = M$ (i.e., stop testing and declare mastery) if $\theta_c < \underline{\theta}_k$; set $K_T = k$ and $D_T = N$ (i.e., stop testing and declare nonmastery) if $\overline{\theta}_k < \theta_c$; or continue testing to stage $k + 1$ if $\underline{\theta}_k \leq \theta_c \leq \overline{\theta}_k$. If $k = n$, the test is necessarily terminated. Set $D_T = M$ if and only if $\theta_c \leq \hat{\theta}^{(n)}$.

The above formulation of the truncated CI stopping rule in AMT is simple and thus can be implemented very easily in practice. To see why further improvement using curtailment is beneficial, consider, for example, the following hypothetical situation: On a test with $n = 6$ and cutoff point $\theta_c = -.2$, suppose that the first five items administered to a certain examinee have discrimination parameters $a_1 = 1.9$, $a_2 = 1.4$, $a_3 = 1.7$, $a_4 = 1.6$, $a_5 = 1.7$; difficulty parameters $b_1 = 0$, $b_2 = 2.0$, $b_3 = 0.8$, $b_4 = 1.5$, $b_5 = 0.9$; and guessing parameters $c_1 = 0.21$, $c_2 = 0.23$, $c_3 = 0.23$, $c_4 = 0.18$, $c_5 = 0.22$. Suppose that the examinee answers Questions 1, 3, and 5 correctly but Questions 2 and 4 incorrectly. In this case, the examinee's ML ability estimate after answering the first five test items is $\hat{\theta}^{(5)} = 1.21$, with corresponding CI for θ given by $[-0.30, 2.72]$. As the CI at this stage still contains the cutoff point, the test cannot be terminated and the last test item needs to be administered, say, with item parameters $a_6 = 2.0$, $b_6 = 1.2$, $c_6 = 0.19$. If the examinee answers this item correctly, the ML ability estimate becomes $\hat{\theta}^{(6)} = 1.48$. If the examinee answers this item incorrectly, the ML ability estimate becomes $\hat{\theta}^{(6)} = 0.85$. Under either scenario, the examinee will pass the test because the final ML ability estimate will exceed the cutoff point. Thus, the same classification decision could have been made earlier after the administration of the fifth item, because administering the sixth item cannot possibly change the final classification decision that the examinee passes the test.

The choice of item parameters in the preceding example was based on the assumption that an ideal item pool was available; that is, for any updated ability estimate, there exists an item with difficulty parameter close to the interim ability estimate. Although this setting might not be realistic in practice, the possibility of shortening AMT when the final classification decision is certain—or nearly certain—should be clear. It is in this regard that the notion of curtailment and stochastic curtailment will be used. Both methods strive to make the same final classification decision as the truncated CI stopping rule. With curtailment, the truncated CI stopping rule in AMT can be modified as follows: Let D_C and K_C be the classification decision and stopping time, respectively, of the curtailed CI stopping rule. Then, for any $k \in \{n_0, n_0 + 1, \ldots, n - 1\}$, set $K_C = k$ and $D_C = M$ (i.e., stop testing and declare mastery) if $\{K_T = k, D_T = M\}$ or $\{\theta_c \leq \hat{\theta}^{(k)}$ and $P(D_T = M | K_T = n, u_1, \ldots, u_k) = 1\}$; set $K_C = k$ and $D_C = N$ (i.e., stop testing and declare nonmastery) if $\{K_T = k, D_T = N\}$ or $\{\hat{\theta}^{(k)} < \theta_c$ and $P(D_T = N | K_T = n, u_1, \ldots, u_k) = 1\}$; or continue testing to stage $k + 1$ otherwise. If $k = n$, the test is necessarily terminated. Set $D_C = M$ if and only if $\theta_c \leq \hat{\theta}^{(n)}$.

As seen from the preceding formulation of the curtailed CI stopping rule, the authors only stop early to declare mastery if the current ability estimate is at or above the cutoff point and the examinee has a probability of 1 to be declared a master at the end of the test. This can be checked in practice by calculating the examinee's final ability estimate if all future responses were to be incorrect. If that final ability estimate is still at or above the cutoff score, the test is stopped, and the current mastery classification decision is preserved. Similarly, the authors only stop early to declare nonmastery if the current ability estimate is below the cutoff point and the examinee has a probability of 1 to be declared a nonmaster at the end of the test. To check this, the final ability estimate if all future responses were to be correct is calculated. If the resulting final ability estimate is still below the cutoff score, then the test is stopped, and the current nonmastery classification decision is preserved.

As explained earlier, when the CI stopping rule is used in AMT, test items administered to an examinee are selected to maximize information at his or her interim ability estimate. In the context of curtailment, the identity of all future items to which the all-correct or all-incorrect responses (alluded to earlier) will be assigned needs to be known. The identity of the next test item can easily be obtained by utilizing the test responses given so far. The identity of all other future test items can be figured out iteratively by first calculating the examinee's ability estimate after the next item is answered correctly (if all-correct future responses are assumed) or incorrectly (if all-incorrect future responses are assumed), then choosing another item that maximizes information at the new ability estimate, and repeating the process. As another alternative, a ''representative set'' of items can be used as surrogates for those unknown future items. When performing curtailment and stochastic curtailment of the TSPRT, Finkelman (2008) suggested that the representative set consist of the $n - k$ items that have the maximum Fisher information at the current ability estimate, among the items that have not been administered to the examinee thus far. However, because no one has studied curtailment and stochastic curtailment of the truncated CI stopping rule (prior to the current study), the question of whether such a representative set would work well for this stopping rule has been open. The simulation results presented in ''Simulation'' section show that the preceding representative set does indeed work well for purposes of this study.

Stochastically Curtailed CI Stopping Rule

Standard Formulation of Stochastic Curtailment

A more aggressive modification to the truncated CI stopping rule in AMT is obtained by using stochastic curtailment. Here, instead of requiring that the probability be equal to one that the same final classification decision is obtained should the test continue, the only requirement is that the probability be at or above a specified threshold. For an examinee currently considered a master (i.e., an examinee whose current maximum likelihood estimate [MLE] is greater than or equal to the cutoff point), his or her probability of remaining a master is required to be at least γ' for stopping to occur. For an examinee currently considered a nonmaster (i.e., an examinee whose current MLE is less than the cutoff point), his or her probability of remaining a nonmaster is required to be at least γ for stopping to occur. Here γ' and γ are constants satisfying $.5 < \gamma \leq 1$ and $.5 < \gamma' \leq 1$. The sequential method can be expressed mathematically as follows: Let D_S and K_S be the classification decision and stopping time, respectively, of the stochastically curtailed CI stopping rule. Then, for any $k \in \{n_0, n_0 + 1, \ldots, n - 1\}$, set $K_S = k$ and $D_S = M$ (i.e., stop testing and declare mastery) if

$$\{K_T = k, D_T = M\} \text{ or } \left\{\theta_c \leq \hat{\theta}^{(k)} \text{and } \min_{\theta \in \Theta} P_\theta(D_T = M | K_T = n, u_1, \ldots, u_k) \geq \gamma'\right\}; \tag{3}$$

set $K_S = k$ and $D_S = N$ (i.e., stop testing and declare nonmastery) if

$$\{K_T = k, D_T = N\} \text{ or } \left\{\hat{\theta}^{(k)} < \theta_c \text{ and } \min_{\theta \in \Theta} P_\theta(D_T = N | K_T = n, u_1, \ldots, u_k) \geq \gamma\right\}; \tag{4}$$

continue testing to stage $k+1$ otherwise. If $k = n$, the test is terminated and $D_S = M$ is set if and only if $\theta_c \leq \hat{\theta}^{(n)}$.

The previous formulation of the stochastically curtailed CI stopping rule is very similar to that of the curtailed CI stopping rule presented earlier. In particular, when $\gamma = \gamma' = 1$, the two stopping rules become equivalent. Nevertheless, the stochastically curtailed CI stopping rule is slightly more complicated. Instead of using hypothetical all-correct or all-incorrect future responses to check whether the probability is one that the same classification decision is made at the end of the test, the probability in Equation 3 or 4 needs to be evaluated at a set Θ of θ values. The set Θ would ideally be a 1-point set consisting of the true ability value, such that the probabilities in Equations 3 and 4 are evaluated under the examinee's true ability. However, the true ability is unknown, and thus, another value (or other values) is used in Θ instead. In the standard formulation of stochastic curtailment that is used in conjunction with the TSPRT as a stopping rule, $\Theta = \{\theta_+, \theta_-\}$ where $\theta_+ = \theta_c + \delta$ and $\theta_- = \theta_c - \delta$ form the end points of the indifference region around θ_c. For early stopping to occur, either the probability statement in Equation 3 or the probability statement in Equation 4 must hold under both θ_+ and θ_-. As the concept of the indifference region does not exist in the framework of AMT with the CI stopping rule, θ_+ and θ_- need to be replaced in Θ by some other values. Thus, several of the variations introduced in Finkelman (2010) were used in the present study, and these are explained next.

Probability Calculations of the Stochastically Curtailed CI Stopping Rule

As a first alternative to evaluate the probability calculations in Equations 3 and 4, the authors consider using $\Theta = \{\hat{\theta}^{(k)}\}$. Upon using the interim ML ability estimate $\hat{\theta}^{(k)}$, information about the examinee's true ability obtained from responses to the items administered so far is utilized. Although $\hat{\theta}^{(k)}$ might not provide an accurate estimate of the true ability for very small k, this will not be problematic in practice when there is a certain minimum test length requirement. At any stage $k \geq n_0$, it is safe to assume that $\hat{\theta}^{(k)}$ is already in the proximity of the true ability, so that using its value in the probability calculations of Equations 3 and 4 will provide a reasonable approximation to the true probability under the true ability value. As k increases, consistency of the ML estimator ensures that $\hat{\theta}^{(k)}$ gets closer to the true ability value (H.-H. Chang & Stout, 1993). Consequently, the approximation becomes more accurate.

As a second alternative, the lower or upper end point of the interim CI can be used for θ in the set Θ. The motivation to use this approach instead of the first one described above relates to the fact that $\hat{\theta}^{(k)}$, as an estimate of θ, still has an inherent uncertainty in it. If $\hat{\theta}^{(k)}$ overestimates θ, the probability in Equation 3 will be overestimated, thus increasing the probability of wrongly invoking early stopping with a mastery classification decision. However, if $\hat{\theta}^{(k)}$ underestimates θ, the probability in Equation 4 will be overestimated, thus increasing the probability of wrongly invoking early stopping with a nonmastery classification decision. To overcome such problems, the probability in Equation 3 can be evaluated at $\Theta=\{\underline{\theta}_k\}$ when $\theta_c \le \hat{\theta}^{(k)}$ and in Equation 4 at $\Theta=\{\bar{\theta}_k\}$ when $\hat{\theta}^{(k)}<\theta_c$. This approach of using $\Theta=\{\underline{\theta}_k\}$ or $\Theta=\{\bar{\theta}_k\}$ can therefore be viewed as a more conservative approach than the earlier approach, whereby $\Theta=\{\hat{\theta}^{(k)}\}$ is used; the ATL of the former is expected to be larger than that of the latter, but the PCD should also be higher.

The authors note that in each of the stopping rules introduced above, the ML ability estimate has been used to construct the end points of all CIs. In particular, in the case of the stochastically curtailed CI stopping rule, the ML estimate of θ is used not only in constructing the CI that serves as a criterion for early stopping but also in evaluating the probabilities in Equations 3 and 4. The rest of this section is devoted to explaining how those probabilities can be evaluated using an alternative to the ML estimate due to the latter's minor drawback explained next.

An Alternative to the MLE: The Modified MLE

At any stage k during the test, the ML estimate $\hat{\theta}^{(k)}$ is obtained by maximizing the likelihood function. This is equivalent to maximizing the logarithm of the likelihood function, which is given by $\log L_k \equiv \log L(u_1, \ldots, u_k|\theta) = \sum_{i=1}^{k}[u_i \log\{P_i(\theta)\} + (1-u_i)\log\{Q_i(\theta)\}]$. By taking the derivative of the log likelihood function with respect to θ, it is seen that $\hat{\theta}^{(k)}$ is the solution to the likelihood equation $\sum_{i=1}^{k}\{u_i - P_i(\theta)\}P_i'(\theta)/P_i(\theta)Q_i(\theta)=0$, where $P_i'(\theta)$ is the first derivative of the item response function in Model 1 with respect to θ.

When the 3PL model is used, the likelihood equation simplifies to $\sum_{i=1}^{k} a_i\{P_i(\theta)-c_i\}\{u_i - P_i(\theta)\}/P_i(\theta)(1-c_i)=0$. Unfortunately, this likelihood function might possess several maxima in $(-\infty,\infty)$, which means that a unique solution to it might not necessarily exist (Hambleton, Swaminathan, & Rogers, 1991). To address the multiple root problem of the likelihood equation, H.-H. Chang and Ying (2009) used the approximation $(P_i(\theta)-c_i)/(P_i(\theta)(1-c_i))\approx(1+\sqrt{1+8c_i})/(2c_i+1+\sqrt{1+8c_i})$ to come up with an approximate likelihood equation:

$$\sum_{i=1}^{k}\frac{a_i\left(1+\sqrt{1+8c_i}\right)\{u_i-P_i(\theta)\}}{2c_i+1+\sqrt{1+8c_i}}=0. \tag{5}$$

The solution to Equation 5, which is called the modified ML estimate of θ, is unique whenever it exists because the left-hand side of 5 is a monotone decreasing function of θ. A necessary and sufficient condition for existence is as follows:

$$\sum_{i=1}^{k}\frac{a_i\left(1+\sqrt{1+8c_i}\right)u_i}{2c_i+1+\sqrt{1+8c_i}}>\sum_{i=1}^{k}\frac{a_i\left(1+\sqrt{1+8c_i}\right)c_i}{2c_i+1+\sqrt{1+8c_i}}. \tag{6}$$

For large n, Equation 6 holds almost certainly, and the corresponding solution $\hat{\theta}_n$ to 5 is strongly consistent as well as asymptotically normal with a mean equal to the true ability and variance $1/FI^{(n)}(\hat{\theta}_n)$. Therefore, a CI of θ can be constructed in the same way as in Equation 2, only replacing the ML estimate of θ by its modified ML estimate. Based on their modified ML estimate of θ, H.-H. Chang and Ying (2009) further proposed the following item selection algorithm, assuming that there exists an infinite item pool and that all item parameters satisfy $0<m<a_i<M<\infty$, $-\infty<b_i<\infty$, and $c_i \leq 1-\delta_0$ for some $\delta_0>0$:

1. Select the first item with parameters a_1, b_1, and c_1. Subsequently, choose items with nondecreasing b parameters if the responses are all correct or items with nonincreasing b parameters if the responses are all incorrect. Continue until the response pattern contains both correct and incorrect answers, say, after k_0 items have been administered.
2. For each $k \geq k_0$, check whether Equation 6 is satisfied. If it is, define $\hat{\theta}_k$ to be the unique solution of Equation 5. Otherwise, define $\hat{\theta}_k = r_k$, where $r_k \downarrow -\infty$ is a predetermined sequence.
3. With $\hat{\theta}_k$ from Step 2, choose the next item such that $b_{k+1} = \hat{\theta}_k$.

In AMT with the curtailed and stochastically curtailed CI stopping rules, the probabilities in Equations 3 and 4 need to be evaluated, conditioning on all item responses that have been observed. Therefore, to use the consistency and asymptotic normality results of H.-H. Chang and Ying (2009) in AMT, a slight generalization is needed for conditional distributions. For this purpose, Theorem 1 will be used, which is available in Appendix A as an online supplement to this article. Using the modified ML estimate of θ proposed by Chang and Ying and Theorem 1, stochastic curtailment for the CI stopping rule in AMT can now proceed by making use of asymptotic theory. Using the result of Theorem 1, the probability calculations in Equations 3 and 4 can now be approximated by a normal probability.

In practice, the asymptotic result based on Theorem 1 is used to decide whether to terminate early only when k is far less than n. If k is close to n (say, with five remaining items on the test), the probability in Equation 3 can be evaluated using exact calculations. This is done by considering all possible response patterns that the examinee can generate for the hypothetical future items from the representative set. With five remaining items on the test before reaching the maximum test length, an exact calculation can be evaluated by considering all $2^5 = 32$ response patterns that the examinee can generate when answering those five items selected from the representative set. A hypothetical final ML ability estimate is then computed in each case, and the proportion of all estimates that meet or exceed the cutoff point is then evaluated, weighted by the respective probability of each response pattern. The probability calculation in Equation 4 is obtained analogously via the response patterns for which the final ML estimate does not reach the cutoff point.

Simulation

The previous section has described motivation for using curtailment as well as stochastic curtailment in conjunction with the truncated CI stopping rule in AMT. In this section, results of a simulation study are presented to provide comparisons between the stopping rules in terms of their ATL as well as PCD. The four stopping rules studied are the original truncated CI stopping rule, the curtailed CI stopping rule, and the two formulations of the stochastically curtailed CI stopping rule described in ‘‘Stochastically Curtailed CI Stopping Rule ’’ section. In the simulation set presented herein, the stopping rules were compared when test items were selected purely based on a psychometric criterion. Another simulation set (available in Appendix B as an online supplement to this article) compared the stopping rules when other test constraints such as content balancing and item exposure control were considered.

Simulation Design

The 3PL model was used with an item pool consisting of 500 items with IRT a-parameters from $U(0.5, 2.5)$, b-parameters from $U(-3.6, 3.6)$, and c-parameters from $U(0.00, 0.25)$ as in H.-H. Chang and Ying (1996). Following Finkelman (2010), $n_0 = 20$ and $n = 50$ were used as the minimum and maximum test length, respectively. All CI-based stopping rules used a 95% CI. In addition, both γ and γ' were fixed at .95 for the stochastically curtailed stopping rule. The cut-off point for the test was $\theta_c = 0$, reflecting a testing situation where approximately half of the examinees would pass under a standard normal ability distribution in the population. One thousand replications were conducted at each of 13 evenly spaced θ values from $\theta_c - 0.6$ to $\theta_c + 0.6$, and results were averaged across replications.

To avoid the multiple roots problem of the likelihood equation of the 3PL model as discussed in the ''An Alternative to the MLE: The Modified MLE'' section, the modified ML estimate of H.-H. Chang and Ying (2009) was used for all four stopping rules. In particular, the truncated CI stopping rule was based on the CI of θ that was constructed at each stage following 2, only with the modified ML estimate of θ used in place of the original ML estimate. The curtailed CI stopping rule was based on computing the final modified ML ability estimate should all future responses be correct (or incorrect) for an examinee whose current modified ML ability estimate is below (or above) the test cut-off point. The stochastically curtailed CI stopping rule was based on the probability calculations in Equations 3 and 4. Two versions were examined: (a) evaluating the probabilities at $\Theta = \{\hat{\theta}_k\}$ and (b) evaluating the probabilities at the lower or upper end point of the interim CI for θ, depending on the relative position of the interim modified ML ability estimate to the test cutoff point. It can be noted that the simulations were also performed using the original ML estimate and the results were similar to those with the modified ML estimate; therefore, for simplicity, only the results using the modified ML estimate are presented. In addition, the simulations using modified ML estimates were performed with two approaches to evaluate the probability in Equation 3: one when it was calculated using exact calculations for $k \in [n-5, n)]$, and the other when the asymptotic result based on Theorem 1 was used throughout the test. While only results under the first approach will be presented hereafter, it was found that there were only minor differences between results under both approaches. Therefore, continuing to use the asymptotic result based on Theorem 1 is also an option in practice.

In all stopping rules and both simulation sets, test items were selected following the algorithm proposed by H.-H. Chang and Ying (2009) as described in ''Stochastically Curtailed CI Stopping Rule'' section. In particular, Equation 6 was checked at each stage during the test. If the condition was met, $\hat{\theta}_k$ was defined to be the unique solution of Equation 5. Otherwise, $\hat{\theta}_k$ was set to be the minimum of $\{\hat{\theta}_1, \ldots, \hat{\theta}_{k-1}, r_k\}$, where $\{r_k\}$ is a decreasing sequence with $r_1 = 0$ and $r_n = -4$. For the curtailed and stochastically curtailed CI stopping rules, the identity of future items was approximated using the notion of ''representative set'' of Finkelman (2008).

Similar to Finkelman (2008, 2010), each simulation set was conducted to allow for a matched comparison between the different stopping rules. For every simulated examinee, the four stopping rules were run simultaneously, using the same response pattern to make a classification decision. If one stopping rule ended before the others, the remaining methods were allowed to continue until they reached their respective stopping time. To evaluate accuracy of classification decisions, a correct decision was defined as mastery if the true ability of the examinee was greater than or equal to the test cutoff point and nonmastery otherwise.

Simulation Results

Table 1 presents the PCD, ATL, as well as average losses (explained in the following) for the four stopping rules being compared.

Comparing the original (truncated) CI stopping rule with its curtailed version, the difference in PCD was never higher than 0.009 = 0.9% at any ability value. In the opening section, it was explained that in theory, the PCD of the curtailed stopping rule would always be the same as that of the original method. The differences observed here between the PCDs of the two methods can be understood by revisiting the notion of ''representative set'' used in the simulation. Due to the adaptive nature of the item selection algorithm, the identity of future items that would be administered should the test continue was unknown. A set of approximate future items was therefore needed for the curtailed stopping rule to compute the final modified ML ability estimate of each examinee under hypothetical all-incorrect (or all-correct) future responses. In reality, the original CI stopping rule might proceed with a different set of future items chosen adaptively based on the examinee's subsequent responses. If the two sets of items are exactly the same, then the PCDs of the two methods will also be the same. This is the case, for example, if test items are chosen based on maximizing Fisher information at the cutoff point as is commonly the case when the TSPRT is used as a stopping rule in AMT. Having seen that the curtailed CI stopping rule did not sacrifice much classification accuracy compared with the original method, it is of interest to see how many test items the former can save. Comparing the ATL of the original CI stopping rule and that of its curtailed version, the latter was able to save an average of 2.62 items across all 13 θ values. More specifically, the latter reduced the ATL by two items or more at 9 of the 13 θ values, by three items or more at 6 values, and by four items or more at 2 values.

Turning to the more aggressive stochastically curtailed CI stopping rule (i.e., the rule whereby the probability calculations in Equations 3 and 4 are evaluated at the modified ML estimate $\hat{\theta}_k$), its PCD was always within 0.033 = 3.3% of that of the original CI stopping rule. The median PCD difference was only 1% in favor of the latter. However, the former reduced the ATL by an average of 11.33 items across all 13 θ values. Indeed, at 7 of the 13 θ values, the stochastically curtailed CI stopping rule was able to save more than 13 items compared with the original stopping rule, and at 5 values, it was able to save more than 15 items.

The more conservative stochastically curtailed CI stopping rule (i.e., the rule whereby the probability calculations in Equations 3 and 4 are evaluated at the lower and upper end point of the CI of θ, respectively) also enhanced efficiency. Its PCD was always within 0.013 = 1.3% of the PCD of the original CI stopping rule. However, it reduced the ATL by an average of 5.61 items across all 13 θ values. More specifically, it saved at least five test items at 9 of the 13 θ values and at least seven items at 6 values.

To further evaluate whether the shorter test lengths of the curtailed and stochastically curtailed stopping rules are worth considering despite their slightly lower PCDs, each of these three methods was compared with the original CI stopping rule using a classification efficiency index proposed by Vos (2000). With 1_W denoting a dummy variable that indicates an incorrect classification, C_W denoting the cost of classification error, and L denoting test length, define

$$\text{Loss} = (1_W \times C_W) + L. \tag{7}$$

This index, which ranks different stopping rules by taking both PCD and ATL into account, has also been used by Finkelman (2008, 2010) in the context of AMT with variants of the TSPRT stopping rule.

Table 1 presents the average loss at each θ value with $C_W = 100$ (Vos, 2000). At all θ values, the more aggressive stochastically curtailed CI stopping rule proved to be the best termination

Table 1. Simulation Set 1: Proportion of Correct Decisions, Average Test Lengths, and Average Losses With $C_w = 100$.

	Truncated CI			Curtailed CI			SC CI modified ML			SC CI with CI end point		
θ	PCD	ATL	Average loss	PCD	ATL	Average loss	PCD	ATL	Average loss	PCD	ATL	Average loss
−.60	1.000	23.41	23.41	1.000	22.86	22.86	0.995	20.35	20.85	1.000	21.82	21.82
−.50	0.995	26.36	26.86	0.995	25.12	25.62	0.986	20.83	22.23	0.994	23.34	23.94
−.40	0.984	31.27	32.87	0.985	28.77	30.27	0.975	21.99	24.49	0.985	25.76	27.26
−.30	0.964	36.73	40.33	0.964	33.21	36.81	0.937	23.69	29.99	0.964	29.51	33.11
−.20	0.873	40.65	53.35	0.876	36.54	48.94	0.875	25.00	37.50	0.877	32.91	45.21
−.10	0.712	44.24	73.04	0.716	40.14	68.54	0.689	28.04	59.14	0.717	36.74	65.04
.00	0.519	46.37	94.47	0.511	42.47	91.37	0.498	28.61	78.81	0.506	38.79	88.19
.10	0.706	44.46	73.86	0.697	41.04	71.34	0.693	28.42	59.12	0.693	37.40	68.10
.20	0.895	42.48	52.98	0.896	38.82	49.22	0.862	25.96	39.76	0.890	34.57	45.57
.30	0.964	37.20	40.80	0.962	34.23	38.03	0.951	23.94	28.84	0.961	30.28	34.18
.40	0.989	32.58	33.68	0.989	30.30	31.40	0.979	22.10	24.20	0.988	27.10	28.30
.50	0.999	27.88	27.98	0.999	26.52	26.62	0.996	20.91	21.31	0.999	24.13	24.23
.60	1.000	23.77	23.77	1.000	23.32	23.32	0.999	20.33	20.43	1.000	22.08	22.08

Note. CI = confidence interval; SC = stochastic curtailment; ML = maximum likelihood; PCD = proportion of correct decisions; ATL = average test length.

criterion, followed by the more conservative approach, the curtailed version, and finally the original CI stopping rule. The second simulation set that was performed with additional test constraints showed the same relative performance of the four stopping rules. In particular, the ATL reduction was greater when test constraints were imposed on the simulated tests, a pattern consistent across all three stopping rules that were compared with the original CI method. This is consistent with the finding of Finkelman (2008) where the TSPRT was used as the termination criterion.

Summary and Discussion

The purpose of this article was to implement curtailment and stochastic curtailment in the context of AMT with the CI stopping rule. The goal of these methods is to provide savings of test items without unduly compromising classification accuracy. Previous research (Finkelman, 2008, 2010) had demonstrated the success of both methods when the TSPRT stopping rule is used. More recently, stochastic curtailment was also applied in conjunction with the GLR stopping rule and was shown to yield increased efficiency while maintaining similar accuracy (Huebner & Fina, 2014). However, as explained in the Introduction, no previous research had applied curtailment or stochastic curtailment to the CI stopping rule, which is preferable to the TSPRT and the GLR in some testing contexts.

The results of the simulation study confirmed the usefulness of the methods in terms of their relative ATLs and PCDs compared with the original CI stopping rule. On a test with a maximum test length of 50 items and no constraints, the more aggressive stochastically curtailed stopping rule reduced the ATL of the original CI stopping rule by an average of more than 11 items without reducing the PCD by more than 3.3%. Based on the loss function in Equation 7 with $C_W = 100$ as in Vos (2000), the more aggressive stochastically curtailed CI stopping rule was more efficient than the original CI method. In particular, the cost of misclassification C_W must be set at 483 for the original CI stopping rule to exhibit better average loss than its stochastically curtailed version at even 1 of the 13 θ values.

The results presented in this article were obtained using the modified ML ability estimate of H.-H. Chang and Ying (2009). This ability estimate can be computed fairly easily by conducting a grid search of the solution to Equation 5 when the inequality in Equation 6 holds. Otherwise, the interim ability estimate will be assigned a value from a decreasing sequence $\{r_n\}$. Using the modified ML ability estimate, the regular CI stopping rule can be applied in the same manner as if the MLE were used, due to the asymptotic equivalence of the two estimators. To invoke stochastic curtailment, the normal approximation based on Theorem 1 provides a convenient way to evaluate the probabilities in Equations 3 and 4. In practice, the regular ML ability estimate can also be used if there is no concern regarding its uniqueness. As mentioned earlier, the simulation study was also conducted using the ML ability estimate and the results were consistent. As an illustration, Figure 1 in Appendix C (available online as a supplement to this article) shows that the ML and the modified ML estimates were generally similar to each other. The figure shows comparisons between both types of estimators for four ability levels (-0.6, -0.2, $+0.2$, and $+0.6$). For each ability level, the result displayed is the average across all 1,000 examinees. As each examinee receives different numbers of items, the results are only displayed up to the 20th item, which is the minimum number of items that each examinee receives.

In this article, the authors focused on evaluating whether curtailment and stochastic curtailment improve the CI stopping rule. With evidence favoring the curtailed and stochastically curtailed CI stopping rules, a possible direction for future research is to compare them with the curtailed and stochastically curtailed versions of the TSPRT and the GLR. In addition, performance of the methods also needs to be further investigated under different structures of item pools, different minimum and maximum test lengths, and other test constraints in addition to content balancing and item exposure control. Extending the methods to classification tests with more than 1 cutoff point also provides fruitful research opportunities, as many statewide tests under No Child Left Behind typically classify students to multiple proficiency groups (Finkelman, 2010).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

The online appendix is available at http://apm.sagepub.com/supplemental

References

Bartroff, J., Finkelman, M., & Lai, T. L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika*, *73*, 473-486.

Chang, H.-H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 37-52.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213-229.

Chang, H.-H., & Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, *37*, 1466-1488.

Chang, Y.-C. I. (2005). Application of sequential interval estimation to adaptive mastery testing. *Psychometrika*, *70*, 685-713.

Eisenberg, B., & Ghosh, B. K. (1980). Curtailed and uniformly most powerful sequential tests. *The Annals of Statistics*, *8*, 1123-1131.

Finkelman, M. D. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics*, *33*, 442-463.

Finkelman, M. D. (2010). Variations on stochastic curtailment in sequential mastery testing. *Applied Psychological Measurement*, *34*, 27-45.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Huebner, A. R., & Fina, A. D. (2014). The stochastically curtailed generalized likelihood ratio: A new termination criterion for variable-length computerized classification tests. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-014-0490-y

Lan, K. K. G., Simon, R., & Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics, Part C: Sequential Analysis*, *1*, 207-219.

Reckase, M. D., & Spray, J. A. (1994, April). *The selection of test items for decision making with a computer adaptive test*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Spray, J. A. (1993). *Multiple-category classification using a sequential probability ratio test* (ACT Research Report Series No. 93–7). Iowa City, IA: American College Testing.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedure for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics*, *21*, 405-414.

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

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research & Evaluation*, *16*(4). Retrieved from http://pareonline.net/getvn.asp?v=16&n=4

Vos, H. J. (2000). A Bayesian procedure in the context of sequential mastery testing. *Psicológica*, *21*, 191-211.

Wald, A. (1947). *Sequential analysis*. New York, NY: John Wiley.

Weiss, D. J. (Ed.). (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York, NY: Academic Press.

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

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement*, *67*, 41-58.