Accuracy and Efficiency in Classifying Examinees Using Computerized Adaptive Tests: An Application to a Large Scale Test

Steven W. Nydick

University of Minnesota

Yuki Nozawa Rongchun Zhu

ACT, Inc.

Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

April 2012

Abstract

Computerized classification testing (CCT) is a modification of computerized adaptive testing (CAT) with the goal of classifying examinees into pre-specified categories. A major component of every classification test is determining at what point a classification decision should be made. One frequently used stopping rule in CCT is the Sequential Probability Ratio Test (SPRT), which results in a classification either when the strength of the log-likelihood ratio test statistic is sufficiently large or when the maximum number of items has been reached. In short tests, the SPRT is inefficient due to properties of the likelihood ratio, necessitating other methods that address shortcomings of the SPRT, including the Generalized Likelihood Ratio (GLR) and the SPRT with Stochastic Curtailment (SCSPRT). The SCSPRT terminates a classification test when the probability of switching categories by maximum test length is small. Because most of the work on stopping rules was derived for a two category CCT, the current study compares the SPRT, GLR, and SCSPRT under a variety of conditions when there are more than two categories. None of the stopping rules adequately control the misclassification rate, but as expected, the SCSPRT results in shorter tests than the SPRT without much loss in classification accuracy. Many practitioners might prefer to use the GLR, which resulted in similar performance to the SCSPRT but with a substantial decrease in computation time.

Accuracy and Efficiency in Classifying Examinees Using Computerized Adaptive Tests: An Application to a Large Scale Test

1. Introduction

Many educational and workforce assessments must classify examinees into pre-specified categories. A basic classification decision might determine whether a particular examinee is proficient at a task based on a single cut score (e.g., Kingsbury & Weiss, 1983; Welch & Frick, 1993; Yang, Poggio, & Glasnapp, 2006) and is frequently referred to as a mastery or certification test. However, finer-grained divisions are often needed, such that classifying an examinee into one of a number of categories can profoundly influence educational success or career satisfaction. For example, classification tasks might be intended to identify student proficiency level related to state standards or a job candidate's skill level compared to job requirements.

Regardless of the intent of classification, writing and administering many items to each examinee can be costly and inefficient. Computerized classification testing (CCT) applies adaptive testing methodology (e.g., Wainer, 2000; Weiss, 1982; Weiss & Kingsbury, 1984) to reduce the number of items administered to a given examinee (e.g., Eggen, 1999; Eggen & Straetmans, 2000; Finkelman, 2008; Jiao & Lau, 2003; Kalohn & Spray, 1999; Lewis & Sheehan, 1990; Thompson, 2007). By using an optimal algorithm to select items and determine the number of items to administer, test practitioners efficiently classify examinees into the appropriate category. Because CCT requires decision rules to choose between competing categories, psychometricans have applied sequential methods taken from statistical decision theory (c.f., Wald, 1947) to adaptive

classification algorithms (e.g., Eggen, 1999; Finkelman, 2003, 2008; Spray & Reckase, 1996; Thompson, 2009; Wouda & Eggen, 2009).

The most commonly used sequential algorithm in CCT, the Sequential Probability Ratio Test (SPRT; Wald, 1947) determines when enough independent and identically distributed data (i.i.d.) have been collected to choose between one of two simple hypotheses (e.g., Keener, 2010, pp. 417-422). With regard to classification testing, these simple hypotheses are quantified as ability levels within each category. The original SPRT procedure has been modified in different ways: (1) Extending the CCT using SPRT to more than two categories (e.g., Eggen, 1999; Spray, 1993); (2) Using composite hypotheses that take into consideration an examinee's current ability estimate (GLR; Thompson, 2009b, 2010); and (3) Making a classification decision when the probability of being classified in the current category by maximum test length is high given the remaining items in the bank (SCSPRT; Finkelman, 2003, 2008; Lan, Simpon, & Halpern, 1982). In this study, we applied the SPRT, GLR, and SCSPRT to a realistic, computerized classification task with three and five categories. By applying various CCT termination methods, we hope to provide corroborative evidence that sequential methods work well in realistic testing scenarios.

The remainder of this paper is organized as follows. In Section 2, we identify the IRT model underlying CCT, and we define the hypotheses needed to make any decision. In Section 3 and 4, we introduce the SPRT and GLR as solutions to the classification problem, and in Section 5, we explain the SCSPRT as an alternative method of classification. In Sections 6 and 7, we outline our current study and describe the results.

Finally, in Section 8, we evaluate the overall simulation, discuss limitations, and propose future directions.

2. Item Response Theory and Computerized Classification Testing

Item Response Theory (IRT) formalizes the relationship between responses to test items and examinee ability. The most common IRT model remains the unidimensional, three-parameter logistic, binary response model (3PL model; Birnbaum, 1968). Let θ denote the ability variable underlying responses to test items, assume that item responses are conditionally independent given a particular level of θ (call that level θ_i for person *i*), and allow all items to have two possible outcomes: a correct and incorrect response. Then, according to the 3PL, the probability of person *i* correctly responding to item *j* can be represented with the following item response function (IRF):

$$p_j(\theta_i) = \Pr(u_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + \exp[-Da_j(\theta_i - b_j)]} \quad (1)$$

where u_{ij} designates the 0/1 response of person *i* to item *j*, b_j is the difficulty or extremity parameter of item *j*, a_j is the discrimination parameter of item *j*, c_j is the lower asymptote for item *j*, and *D* is a scaling constant: 1.00 for the logistic metric and 1.702 for the normal-ogive metric. Furthermore, because *D* is a scaling constant that does not affect model fit, we subsequently absorb *D* into the discrimination parameter for clarity.

By defining ability as a single, latent variable falling along a continuum, we can easily systematize the classification problem. For instance, take a mastery test with two categories, and denote the ability level separating a master from a non-master as θ_0. True classification depends on where an actual examinee falls in relation to the cutting point, θ_0. If $\theta_i \geq \theta_0$ for examinee *i*, then the examinee falls into category 1 and should pass the

test; alternatively, if $\theta_i < \theta_0$, then the examinee falls into category 0 and should fail the test. However, we never know true θ_i, so we have to make a decision with incomplete information. Following Finkelman (2008), label the decision made for person *i*, D_i. If $\theta_i \geq \theta_0$ and $D_i = 1$ or $\theta_i < \theta_0$ and $D_i = 0$, then we have made a correct decision. Otherwise, we are in error. Sequential tests were derived to use the fewest items while strictly controlling the error rate.

3. The SPRT and TSPRT as Applied to CCT

The classic SPRT in classification CAT (e.g., Eggen, 1999; Reckase, 1983; Spray & Reckase, 1986) starts out by defining a simpler form of the classification problem. Rather than testing whether an examinee is above a cut-point versus below a cut-point, the hypotheses are simplified to the end points of an indifference region surrounding the cut-point. If the indifference region is symmetric, then we can denote the indifference region as ($\theta_0 - \delta$, $\theta_0 + \delta$). The indifference region should be set such that any true ability within it can be classified without regret as either passing or failing the test.

After an examinee responds to an item, the SPRT calculates the likelihood if his or her actual ability were $\theta_0 + \delta$ (at one end of the indifference region) versus the likelihood if his or her ability were $\theta_0 - \delta$ (at the opposite end of the indifference region). If responses are conditionally independent and follow a unidimensional, binary, item response function (IRF), then the log-likelihood of a particular response pattern, $\mathbf{u}_i$, assuming a true ability of θ_i is

$$\log[L(\theta_i|\mathbf{u}_i)] = \sum_{j=1}^{J} \left[u_{ij} \log[p_j(\theta_i)] + (1 - u_{ij}) \log[1 - p_j(\theta_i)]\right] \quad (2)$$

where $P_j(\theta_i)$ is defined in Equation (1). The log-likelihood of examinee *i* having ability $\theta_u = \theta_0 + \delta$ relative to having ability $\theta_l = \theta_0 - \delta$ is

$$\log\left[LR(\theta_u;\theta_l|\mathbf{u}_i)\right] = \log\left[\frac{L(\theta_u|\mathbf{u}_i)}{L(\theta_l|\mathbf{u}_i)}\right] = \log\left[L(\theta_u|\mathbf{u}_i)\right] - \log\left[L(\theta_l|\mathbf{u}_i)\right] \quad (3)$$

with θ_u and θ_l replacing θ_i on the right-hand side of Equation (2). When Equation (3) is a large, positive number, then there is sizable evidence supporting θ_u as generating the particular response pattern rather than θ_l. Conversely, when Equation (3) is a large, negative number, then there is sizable evidence supporting θ_l.

To determine the degree of evidence that one would need before making a decision, Wald (1947) recommended using $C_l = \log[\beta/(1-\alpha)]$ and $C_u = \log[(1-\beta)/\alpha]$] as lower and upper critical values, where α and β represent the desired Type I and Type II error rates, respectively. When $\log\left[LR(\theta_u;\theta_l|\mathbf{u}_i)\right]$ is below the lower critical value, the CCT should terminate and classify examinee *i* in the lower category. When $\log\left[LR(\theta_u;\theta_l|\mathbf{u}_i)\right]$ is above the upper critical value, the CCT should classify examinee *i* in the upper category. Otherwise, there is not enough evidence for classification, and the examinee should be administered another item. Even though the SPRT tests simple hypotheses, θ_l versus θ_u, when the true value of θ_i is outside of the indifference region, the actual error rates are smaller than the nominal error rates at the end points (see Chang, 2004, p. 49).

Unfortunately, researchers have identified at least two limitations of the classic SPRT in adaptive testing. First, practitioners are seldom interested only in the endpoints of an indifference region. Classification tasks are usually used to determine whether the

examinees evince any trait level between two cut-points. In light of this concern, the Generalized Likelihood Ratio (GLR; Bartroff, Finkelman, & Lai, 2008; Thompson, 2009, 2010) was proposed as a modification of the SPRT for testing composite hypotheses. Second, most classification tasks must decide on a category before some maximum number of items have been administered, *K*. After *K* items, the test forces termination by comparing the log-likelihood ratio to the average of the two critical values, C_l and C_u. Even though the original SPRT is thought to have optimal properties, in that "when observations are independent and identically distributed (IID), no other method has better error rates and better average test lengths" (Finkelman, 2008, p. 451), Finkelman (2003, 2008) demonstrated that a truncated SPRT is not the most efficient decision procedure. In the next two sections, we elaborate on adjustments to the SPRT algorithm.

4. The GLR

The Generalized Likelihood Ratio (GLR) is a modification of the original SPRT algorithm for testing composite hypotheses. The original SPRT hypotheses are as follows (for a given classification bound θ_0):

$$H_0: \theta_i = \theta_0 - \delta$$
$$H_1: \theta_i = \theta_0 + \delta$$

These hypotheses are the same for the *entire test* regardless of the shape of the likelihood function. Bartroff, Finkelman, and Lai (2008) and Thompson (2009b, 2010) proposed a method by which the algorithm searches for the maximum of the likelihood function on either side of the boundary point and compares that maximum with the threshold on the other side (see Thompson, 2010, p. 5 for a graphical representation of the modified test statistic). Importantly, because the hypotheses can vary as a function of the maximum likelihood estimate ($\hat{\theta}_i$), the modified GLR procedure is effectively testing

$$H_0: \theta_i \leq \theta_0 - \delta$$
$$H_1: \theta_i \geq \theta_0 + \delta$$

which are the original hypotheses rather than the simplified ones used to calculate the SPRT likelihood ratio statistic. Batroff et al. (2008) recommended basing part of the test statistic (including the critical values) on simulation, whereas Thompson (2009, 2010) suggested using the same critical values as in the SPRT. Even when using an inexact system, Thompson (2010) found that the classification accuracy when using the GLR was not noticeably different than the SPRT but that the average test length was reduced by approximately 10 items on a K=200 item test.

Although the GLR has been shown to outperform the SPRT, neither procedure is optimal unless the maximum test length is unlimited. Finkelman (2003, 2008) showed that for $K < \infty$, a procedure that terminated a classification test when the probability of switching categories by maximum test length was small would outperform the SPRT given the same α, β, and δ. In the next section, we describe the logic of stochastic curtailment as applied to classification CAT.

5. The SPRT with Stochastic Curtailment

Finkelman (2004, 2008) developed a supplementary classification criterion, based on Lan, Simon, and Halpern (1982), which he termed the Sequential Probability Ratio Test with Stochastic Curtailment (SCSPRT). After administering the maximum number of items, the Truncated SPRT (TSPRT) will be *forced* to make a decision, usually based on the category in which $\hat{\theta}_i$ falls. Even though the likelihood ratio might not satisfy the SPRT criterion by $k < K$, a different classification might be unlikely by $k = K$ due to a

weak set of remaining items. The SCSPRT proceeds in three steps: (1) Sequential Probability Ratio Test, (2) Curtailment, and (3) Stochastic Curtailment.

5.1 Curtailment

We can redefine the curtailment problem in terms of the log-likelihood, as defined in Equation (2). The maximum likelihood estimator of $\hat{\theta}_i$ is determined by setting the derivative of the log-likelihood equal to 0 and solving for θ_i. For the 3PL, the equation simplifies to

$$\sum_{j=1}^{k}[a_j p_j^*(\theta_i)] = \sum_{j=1}^{k} a_j u_{ij}\left(\frac{p_j^*(\theta_i)}{p_j(\theta_i)}\right) \tag{4}$$

where

$$p_j^*(\theta_i) = \frac{1}{1+\exp[-a_j(\theta_i - b_j)]}. \tag{5}$$

Note that we can use Equation (4) to iteratively find a maximum likelihood estimate given a provisional series of responses. Whether a classification test should be curtailed depends on whether particular values of θ have the possibility of being a maximum likelihood estimate. For instance, let the maximum likelihood estimate be less than the cut-point after $k < K$ items. Then if Equation (4) cannot hold for θ_i above the cut-point after the maximum number of items have been administered, then any θ in the upper category will never be the maximum likelihood estimate.

Next, set

$$T_K = \sum_{j=1}^{K}[a_j p_j^*(\theta_0)] \tag{6}$$

and

$$S_k = \sum_{j=1}^{k} a_j u_{ij}\left(\frac{p_j^*(\theta_i)}{p_j(\theta_i)}\right) \tag{7}$$

where $\tilde{\theta}_i$ is the endpoint closest to θ_0 of a confidence interval (described in Section 5.3) and T_K is calculated by combining items administered so far with potential future items remaining in the bank. Then a classification test should terminate if

$$T_K > S_k + \sum_{j=k+1}^{K} a_j \geq S_k + \sum_{j=k+1}^{K} a_j \left(\frac{p_j^{-}(\tilde{\theta}_i)}{p_j(\tilde{\theta}_i)} \right) \tag{8}$$

or if

$$T_K < S_k . \tag{9}$$

When Equation (8) holds, examinee *i* can answer the remaining items correctly without $\hat{\theta}_i > \theta_0$, and when Equation (9) holds, examinee *i* can answer the remaining items incorrectly without $\hat{\theta}_i < \theta_0$. In both cases, additional information will not change the classification.

5.2 Stochastic Curtailment

The "curtailment condition", as defined in Section 5.1, requires the hypothetical examinee to not change categories by the end of the test. Finkelman (2003) added a probabilistic component to the "curtailment condition," so that the test will terminate if the examinee *should* not change categories given the remaining test items. Assume that enough items are remaining so that $S_K | u_i^{(k)}$ (the distribution of S_K at the end of the test given the items already taken) is approximately normally distributed. Then Finkelman (2003) demonstrated that the probability that examinee *i* will be above the cut-point at maximum test length given an MLE currently above the cut-point ($\Pr(\hat{\theta}_{iK} | \hat{\theta}_{ik}) \geq \theta_0$ where $\hat{\theta}_{ik} > \theta_0$) can be written as

$$\Phi\left(\frac{\hat{E}(S_K | u_i^{(k)}) - T_K}{\sqrt{\hat{V}ar(S_K | u_i^{(k)})}} \right). \tag{10}$$

Furthermore, the probability that examinee *i* will be below the cut-point at maximum test length given an MLE currently below the cut-point ($\Pr(\hat{\theta}_{iK}|\hat{\theta}_{ik}) \leq \theta_0$ where $\hat{\theta}_{ik} < \theta_0$) can be written as

$$\Phi\left(\frac{T_K - \hat{E}(S_K|\mathbf{u}_i^{(k)})}{\sqrt{\hat{V}\widehat{ar}(S_K|\mathbf{u}_i^{(k)})}}\right) \quad (11)$$

where

$$E[S_K|\mathbf{u}_i^{(k)}] = S_k + \sum_{j=k+1}^{K}[a_j \bar{p}_j(\theta_i)] \quad (12)$$

and

$$Var[S_K|\mathbf{u}_i^{(k)}] = \sum_{j=k+1}^{K} a_j^2 [p_j \bar{q}_j(\theta_i)]^2 \left(\frac{1 - p_j(\theta_i)}{p_j(\theta_i)}\right). \quad (13)$$

To find the estimated version of Equations (12) and (13), replace true θ_i with $\tilde{\theta}_i$ as defined in Section 5.1 based on the current estimate of θ_i.

When there are only two categories, the "stochastic curtailment" procedure is as follows: (1) set a nominal error rate (where $\gamma < .50$, usually $\gamma = .05$); (2) calculate Equation (10) (if $\hat{\theta}_{ik} > \theta_0$) or Equation (11) (if $\hat{\theta}_{ik} < \theta_0$); (3) terminate the test and set the final category equal to the current category if the cumulative density is greater than some pre-specified $1 - \gamma$.

5.3 SCSPRT with Many Categories

A logical adjustment to any stopping rule when there are more than two categories (based off of work by Sobel & Wald, 1949) is to check the stopping rule at each bound and classify examinee *i* into category *g* if there is evidence that he or she is above category *g* – 1 and below category *g* + 1 (e.g., Eggen, 1999; Spray, 1993). When using the SPRT or GLR, all of these tests can be performed without additional

information. Unfortunately, when using stochastic curtailment "both additional stopping rules need information of items that could be administered in the future … in order to be able to assign examinees to a certain level" (Wouda & Eggen, 2009, p. 6). Both T_K and S_K depend not *just* on the first *k* items administered to an examinee (which are already known prior to calculating those terms), but they need information from the remaining *K* – *k* items that the examinee will take in the hypothetical future. Wouda and Eggen (2009) determined T_{K-k} and $E[S_{K-k}|u_i^{(k)}] = \sum_{j=k+1}^{K}[a_j p_j(\theta_i)]$ based on the ordering items that maximize Fisher information around particular locations. Fisher information for item *j* only depends on a particular value of θ and not the item responses (Lord, 1980). For instance, Wouda and Eggen (2009) chose items to calculate T_{K-k} and $E[S_{K-k}|u_i^{(k)}] = \sum_{j=k+1}^{K}[a_j p_j(\theta_i)]$ based on the endpoint of the confidence interval closest to a particular bound. To calculate the confidence interval bound, one would only need to use the observed Fisher information for the items already administered, or

$$\tilde{\theta}_i = \hat{\theta}_i - \frac{z_{1-\alpha/2}}{\sqrt{\sum_{j=1}^{k} \hat{\mathcal{I}}_j(\hat{\theta}_i)}} \quad \text{or} \quad \tilde{\theta}_i = \hat{\theta}_i + \frac{z_{1-\alpha/2}}{\sqrt{\sum_{j=1}^{k} \hat{\mathcal{I}}_j(\hat{\theta}_i)}} \tag{15}$$

where

$$\hat{\mathcal{I}}_j(\hat{\theta}_i) = -\frac{\partial^2 \log[L(\hat{\theta}_i|u_{ij})]}{\partial \theta_i^2} \quad . \tag{16}$$

Observed Fisher information is slightly different from expected Fisher information for the 3PL model. To find expected Fisher information given a particular value of θ_i, which is needed when choosing items, take the expectation of Equation (16).

Generalizing two classification bounds to *G* classification bounds is relatively simple when adopting the methodology of Spray (1993). With *G* classification bounds, there are *G* + 1 categories and $\binom{G+1}{2}$ possible hypotheses to test. But we only need to

test sequential pairs of hypotheses because, for example, deciding that $\theta_i < \theta_2$ also tells us that θ_i is less than bounds $\theta_3, \theta_4, \cdots, \theta_G$. Therefore, one would only need to test each pair of sequential bounds using the specified stopping rule, and classify a particular examinee into the first category in which the test terminated.

6. Current Study Design

In this section, we describe a simulation study designed to compare stopping rules under a variety of conditions when there are multiple classification categories.

6.1 Assessment Properties

We employed 600 items from a real item bank for a large-scale test calibrated under the assumption of a three-parameter logistic model with D =1.702. The *a*-parameters had a mean of 1.20 and a standard deviation of 0.33; the *b*-parameters had a mean of 0.06 and a standard deviation of 1.43, and the *c*-parameters had a mean of 0.15 and a standard deviation of 0.07. The classification bounds were set to: (1) θ_1 = -0.47 and θ_2 = 1.18 for a three category test, and (2) θ_1 = -1.39, θ_2 = -0.47, θ_3 = 0.28, and θ_4 = 1.18 for a five category test.

There were also minimal content constraints. Items were classified into eight content areas, the first eight items of any CAT were selected to come from each of the eight content areas, and no two consecutive items could come from the same content area.

We repeated the simulation twice using $n_{\min} = 8$ and $n_{\max} = 21$ as the minimum and maximum test length for any simulee. First, we determined the accuracy and test length by simulating 400 classification tests at each of 81 θs evenly spaced

between -4 and 4. We then attempted to mimic the classification accuracy and test length in aggregate by simulating $N = 5000$ θs from a standard normal distribution.

6.2 Item Selection

Regardless of condition, the first four items for any CAT were randomly selected from a range between -0.5 and 0.5, provided that they satisfied the minimal content constraints. Simulees who did not have a mixed response pattern following the first four items were randomly administered very easy or very difficult items until a bounded, maximum likelihood estimate could be calculated. We then implemented two sets of item-selection conditions with various specifications per condition. The "Fisher Information" conditions ordered items according to their expected Fisher information at a particular θ level, and the "Kullback-Leibler Divergence" conditions ordered items according to their KL divergence at a particular θ level. Maximizing Fisher information at a particular ability value is equivalent to minimizing the associated standard error. Kullback-Leibler divergence (Chang & Ying, 1996; Kullback, 1959; Kullback & Leibler, 1951) relates to the expected loss when choosing an approximate model rather than the correct model. For the 3PL IRT model, KL information can be expressed as (Chang & Ying, 1996, p. 217)

$$\begin{aligned} KL_j(\theta_{0+}||\theta_{0-}) &= E_{\theta_{0+}}\left[\log[LR_j(\theta_{0+};\theta_{0-}|\mathbf{u}_i]\right] \\ &= p_j(\theta_{0+})\log\left[\frac{p_j(\theta_{0+})}{p_j(\theta_{0-})}\right] + [1-p_j(\theta_{0+})]\log\left[\frac{1-p_j(\theta_{0+})}{1-p_j(\theta_{0-})}\right]. \end{aligned} \quad (17)$$

Chang and Ying (1996) recommended calculating KL divergence for a particular item by integrating Equation (17) along θ_{0-} within a small distance from θ_{0+}. However, many

practitioners (including Eggen, 1999) estimate KL divergence by setting $\theta_{0+} = \theta_0 + \delta$ and $\theta_{0-} = \theta_0 - \delta$, where θ_0 is the particular ability at which KL divergence should be maximized, and δ is the half-width of the indifference region, as defined in Section 3. Intuitively, an item with the largest KL divergence maximally differentiates between θ slightly larger than θ_0 and θ slightly smaller than θ_0. Fisher information is similar to KL divergence when δ approaches 0.

For the Fisher information and KL divergence conditions, we chose items to maximize the respective criterion at two values of θ: $\hat{\theta}_i$ and the bound closest to $\hat{\theta}_i$. Spray and Reckase (1994) indicated that selecting items to maximize information at the "decision point" (p. 5) performs slightly better than maximizing information at the current estimation of $\hat{\theta}_i$. We chose the closest bound based on Eggen (1999) Method F4. Regardless of item selection method, we chose the top *g* items that satisfied a particular criterion, where *g* was either 1, 5, or 15, and we randomly selected the next item from that set.

6.3 Item Exposure

Due to security concerns, it was desired to limit the proportion of examinees who responded to any given item. We used the Sympson-Hetter (SH; 1985) method of controlling item exposure. The idea behind Sympson-Hetter (e.g., Chen & Lei, 2005) is to control the rate at which examinees see a particular item by limiting the percentage of times an item is administered given that it is selected. To calibrate this probability, we used 5000 simulees from a standard normal distribution, and repeated the entire SH procedure 10 times to stabilize our estimates of the required probabilities.

During CAT administration, when selecting an item for any examinee, a random, uniform variate, *u*, is compared to the probability of an item being administered given that it is selected, $\Pr(A|S)$. If $u < \Pr(A|S)$ the item is administered, but if $u > \Pr(A|S)$, the item is not administered and subsequently removed from the item bank for that examinee. We used two values for the maximum proportion of simulees that should be administered a given item: 1 (no item exposure), and .2 (at most $1/5^{th}$ of the simulees should see any item).

6.4 Ability Estimation

Once a simulee has a mixed response pattern, several item selection and termination procedures require a provisional estimate of θ_i. We used two methods to estimate simulee ability. The first estimation method was via a Maximum Likelihood (MLE) criterion. To determine the MLE estimate, one only needs to iteratively find the maximum of the likelihood equation.

However, Lord (1984) found that the maximum likelihood estimate is biased for fixed length tests, with the bias inversely proportional to the length of the test (see Warm, 1989, p. 428). Warm (1989) presented a correction for the bias inherent in the MLE, which he called Weighted Likelihood Estimation (WLE). Even though many studies have used MLE to estimate ability (e.g., Bartroff, Finkelman, & Lai, 2008; Finkelman, 2008), weighted likelihood estimation has gained a foothold in the classification literature (e.g., Eggen and Straetmans, 2000; Wouda and Eggen, 2009). Moreover, SCSPRT depends on precise estimates of ability to calculate the necessary probabilities, so it is conceivable that WLE could improve the classification accuracy of stochastic curtailment early in a test.

6.4 Termination

Our goal was to compare the classification accuracy and test length of each stopping rule under a variety of conditions. We used four termination conditions, including SPRT, GLR, SCSPRT (each discussed earlier), and Confidence Interval (CI). The last termination condition, CI, is similar to the Sequential Bayes (SB) termination condition discussed in Spray and Reckase (1996). The confidence interval method calculates the confidence interval of the ability estimate, as defined by Equation (17), and classifies a particular examinee into a category when the confidence interval is located solely within that category. Unlike SPRT methods, CI (or SB) does not take into consideration repeatedly estimating θ_i for any examinee, so the specified confidence level overstates the true coverage rate.

7. Simulation Methods and Results

We performed three simulations, each designed to answer different research questions. Our first two simulations examined the classification accuracy and test length conditional on particular levels of ability. The first simulation used a three category classification task, whereas the second simulation used five categories. Our final simulation examined the overall classification accuracy and test length by aggregating over a distribution of ability levels.

7.1 Simulation 1

Methods

We first simulated the conditional test length and classification accuracy for a three category classification test. The specific item selection, ability estimation, and exposure control conditions that we used were described in the previous section. Twelve

stopping rules were used, including: (I) SPRT with δ = .10 or δ = .20 and α = β = .05 or α = β = .10; (II) GLR with δ = .10 or δ = .20 and α = β = .05 or α = β = .10; (III) SCSPRT with $\tilde{\theta}_i$ estimated from a confidence level of 0, 67%, or 95%; and (IV) CI using a confidence level of 95%.

At each combination of conditions, we set true θ to 81, evenly spaced levels, from -4 to 4, and we simulated 400 CATs at each of those levels. Our primary interest was to determine conditional accuracy, number of items, and item exposure at specific θs.

Results

Figure 1 presents the conditional test length for certain termination conditions when items were selected using maximum Fisher information at $\hat{\theta}_i$, ability was estimated using MLE, and there were no item exposure constraints. Consider the green and dark blue curves, which represent SPRT and GLR with identical indifference regions. Although the difference between test lengths for both stopping rules are similar at the classification bounds and for small values of θ_i, the similarity does not extend to θ_i above the highest cut-point. Unlike the SCSPRT, GLR, and CI stopping rules, when using the SPRT as a stopping rule, the increase in efficiency is greater for examinees below the lowest cut-point than above the highest cut-point. The asymmetric efficiency is most likely due to properties of the log-likelihood ratio test statistic in the three-parameter logistic model and explored in a follow-up study.

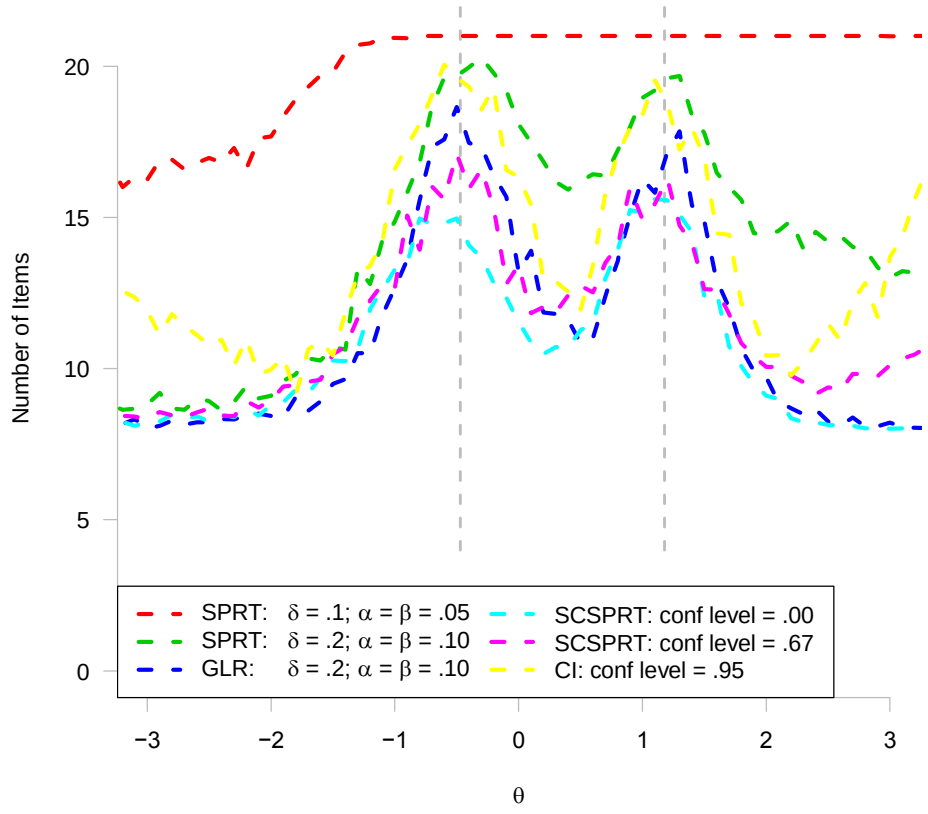

Figure 1

Test length averaged over 400 classification CATs conditional on selected values of θ_i with items selected by Fisher information at $\hat{\theta}_i$, ability estimated by maximum likelihood estimation, and no item exposure control. The vertical bars represent the classification bounds. Only a few termination conditions are presented for illustration purposes.

When comparing the GLR to the SCSPRT conditions, it appears as though stochastic curtailment without a confidence interval correction results in the shortest tests for most levels of ability, and stochastic curtailment with a slight correction results in short tests for true ability near the classification bounds but much longer tests for true ability further away. GLR, on the other hand, results in test lengths about halfway between the SPRT and SCSPRT conditions for true ability close to the classification

bounds but very short tests for true ability in the extremes. A final thing to note is that when terminating tests based on either the confidence interval method or stochastic curtailment with a confidence interval correction, the average test length slopes upward at extreme values of θ_i. Examinees with extreme θ_i tend to answer the first few items in the same direction, so it is impossible to construct a finite confidence interval based on likelihood theory. The increased test length for high and low ability simulees is an artifact of the item bank and would rarely happen if the confidence interval was constructed by Bayesian methods.

Figure 2 displays the conditional accuracy corresponding to the conditions discussed in Figure 1. In Figure 2, all of the stopping rules appear to classify examinees with relatively equal precision for most levels of true ability. However, for those simulees with average or low ability, stochastic curtailment without a confidence interval correction results in the poorest classification accuracy. Note that stochastic curtailment was expected to result in slightly poorer classification accuracy, and the differences between the curves in Figure 2 are small.

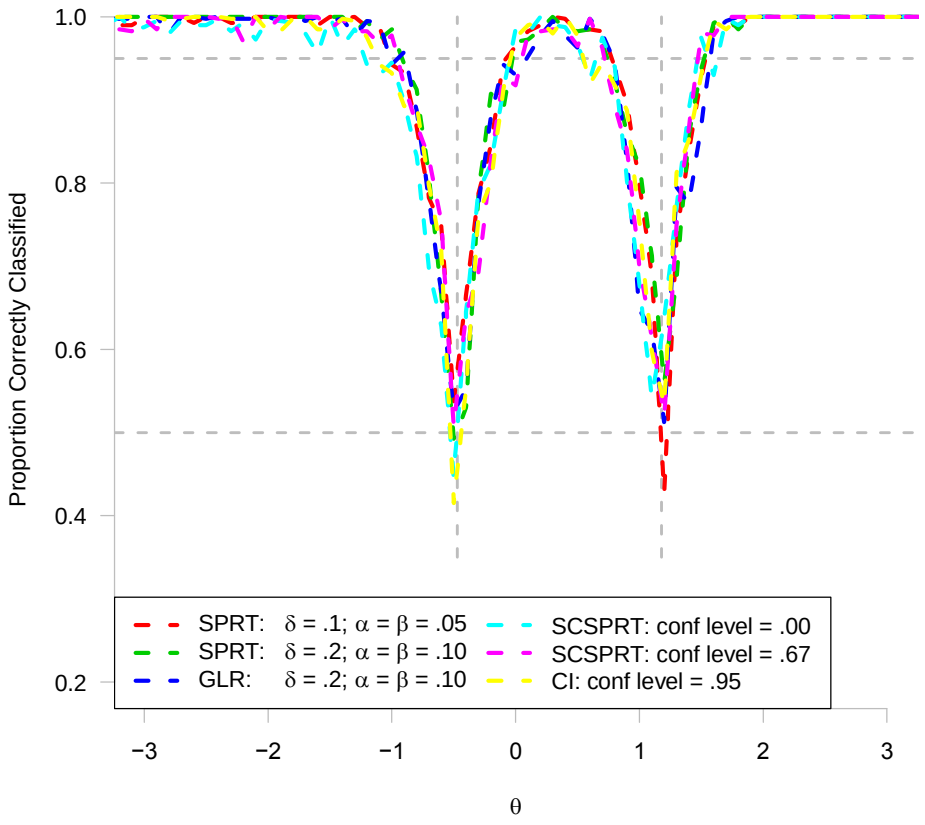

Figure 2

Classification accuracy averaged over 400 classification CATs conditional on selected values of θ_i with items selected by Fisher information at $\hat{\theta}_i$, ability estimated by maximum likelihood estimation, and no item exposure control. The vertical bars represent the classification bounds, and the horizontal bars 50% classification accuracy and 95% classification accuracy. Only a few termination conditions are presented for illustration purposes.

Other combinations of conditions resulted in differences between the stopping rules that were very similar to those presented in Figures 1 and 2. The effect of other factors on test length and classification accuracy will be presented later.

7.2 Simulation 2

Methods

We next applied the procedures described in the previous sections to classify simulees into five categories. The conditions that we used, and the procedure that we followed, were identical to Method 1.

Results

Figure 3 presents the average test length for those combinations of conditions discussed in Figure 1 when there were four cut-points. Notice how with five classification categories, the differences between the stopping rules are magnified as compared to three classification categories. Both stochastic curtailment conditions resulted in the shortest tests for θ_i close to the classification bounds, both SPRT conditions resulted in more efficient tests for θ_i below the lowest cut-point than above the highest cut-point, and the GLR condition resulted in more efficient tests for θ_i in the extremes. The major difference between Figures 1 and 3 is that the relative advantage of stochastic curtailment is magnified when there are multiple cut-points in close proximity.

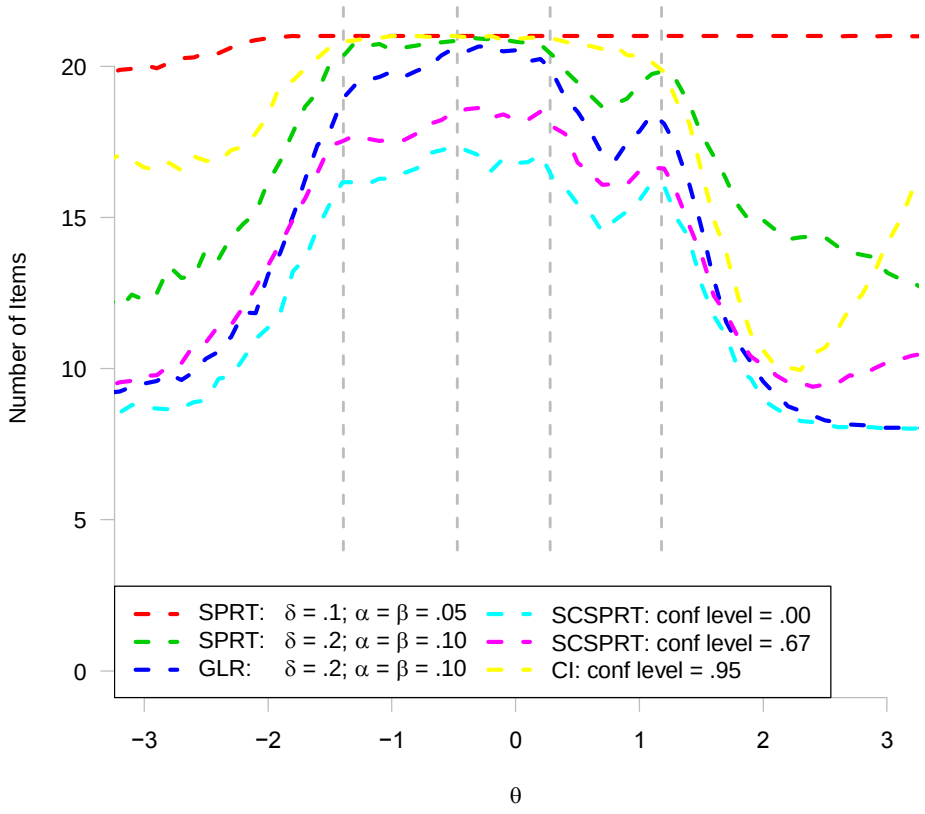

Figure 3

Test length averaged over 400 classification CATs conditional on selected values of θ_i with items selected by Fisher information at $\hat{\theta}_i$, ability estimated by maximum likelihood estimation, and no item exposure control. The vertical bars represent the classification bounds. Only a few termination conditions are presented for illustration purposes.

Figure 4 presents the conditional classification accuracy corresponding to the conditions discussed in Figure 3. Even though using SCSPRT as a stopping rule results in classification tests with the poorest classification accuracy, the differences in accuracy between the stopping rules is less noticeable when there are five categories versus three categories. In conjunction, Figures 3 and 4 suggest that termination criteria resulting in shorter tests do not greatly impact classification accuracy when there are many categories

and a short maximum test length. Furthermore, the relatively small differences in test length for the different stopping rules might add result in different average test lengths up across a realistic distribution of simulees.

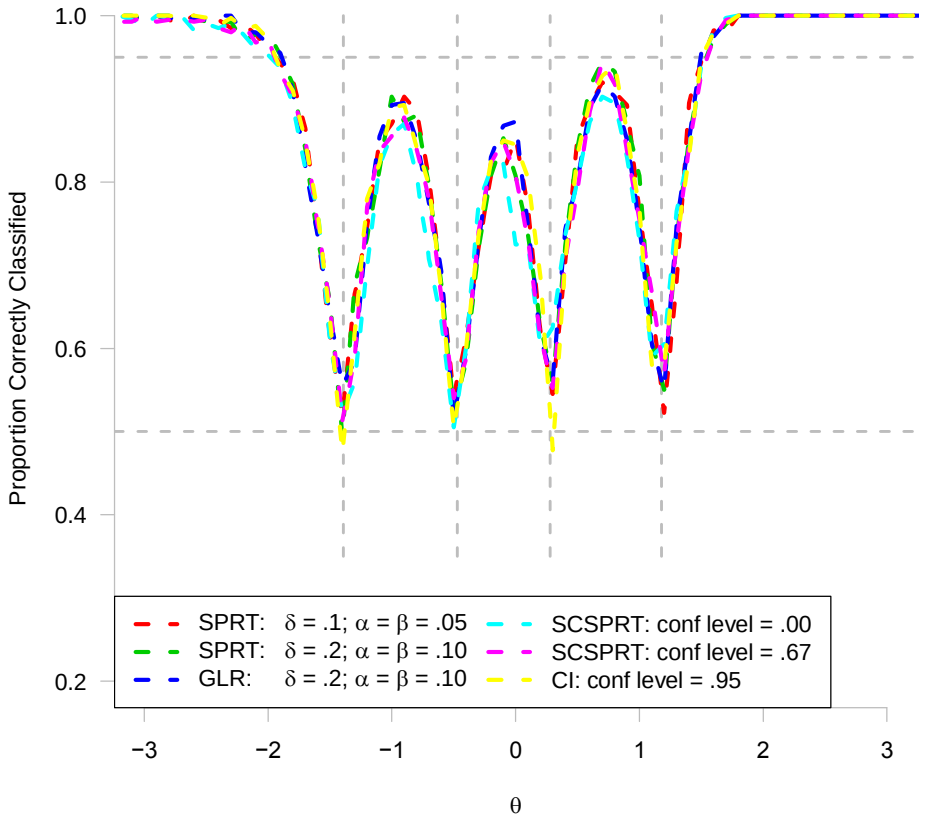

Figure 4

Classification accuracy averaged over 400 classification CATs conditional on selected values of θ_i with items selected by Fisher information at $\hat{\theta}_i$, ability estimated by maximum likelihood estimation, and no item exposure control. The vertical bars represent the classification bounds, and the horizontal bars 50% classification accuracy and 95% classification accuracy. Only a few termination conditions are presented for illustration purposes.

7.3 Simulation 3

Methods

Our final method was designed to estimate the performance of the termination procedures under realistic testing conditions. Rather than fixing ability at 400 evenly spaced values between -4 and 4, we simulated 2500 θs from a standard normal distribution. Only by generating simulees from a distribution could we estimate the overall classification accuracy and test length for a random sample of examinees. The termination, item selection, exposure control, and ability estimation conditions were identical to the previous two methods.

Results

Figure 5 depicts the percent classified correctly and average test length for each combination of conditions in the three category classification task with points shaded to represent different termination conditions. Each of the points within a given color indicates a particular combination of item selection, exposure control, and ability estimation methods. SCSPRT without a correction (represented by the cyan scatter of points) results in the shortest tests. However, the average classification accuracy of the aforementioned stopping rule is worse than any of the other methods by nearly .01. Using SCSPRT with a confidence interval correction of .67 (represented by the pink scatter of points) results in slightly improved accuracy as compared to no correction but with an average of 1 additional item. The most liberal GLR, using $\alpha = \beta = .10$ and $\delta = .20$ (represented by the dark blue scatter of points), results in test lengths and accuracy nearly equivalent to the second SCSPRT condition. Even though there is a slight increase in accuracy when using the standard SPRT, the increase is possibly due to much

longer test lengths. A numerical summary of Figure 5 is displayed in Table 1. The numerical summary displays the first quartile and mean of test length and the third quartile and mean of percent classified correctly for the same conditions. There does not appear to be a strong relationship between test length and percent classified correctly *within termination condition.* As is apparent both in the table and in the graph, there is not much of a difference between the first quartile and the mean of test length with the exception of the confidence interval condition. Moreover, the GLR (with $\alpha = \beta = .10$ and $\delta = .20$) and the SCSPRT (with a conservative correction of .67) results in anywhere from 3–7 items fewer relative to other termination criteria, but with less than a percentage point decrease in accuracy.

Figure 6 displays the corresponding classification accuracy and test length for the five category condition. Unlike the plot shown in Figure 5, there is little noticeable upward trend in accuracy for stopping rules resulting in increased test length. SCSPRT with a conservative correction of .67 appears to be nearly as accurate in classification as GLR and SPRT, but results in anywhere from 1.5–4 items fewer per test. An equivalent numerical summary is shown in Table 2. Note that the third quartile of classification accuracy for the SCSPRT (with a conservative correction of .67) is higher than all but the third quartiles of two SPRT conditions. Both of the SPRT conditions are associated with test lengths close to the maximum allowed (21), whereas the SCSPRT with a conservative correction is associated with a much shorter test length (17).

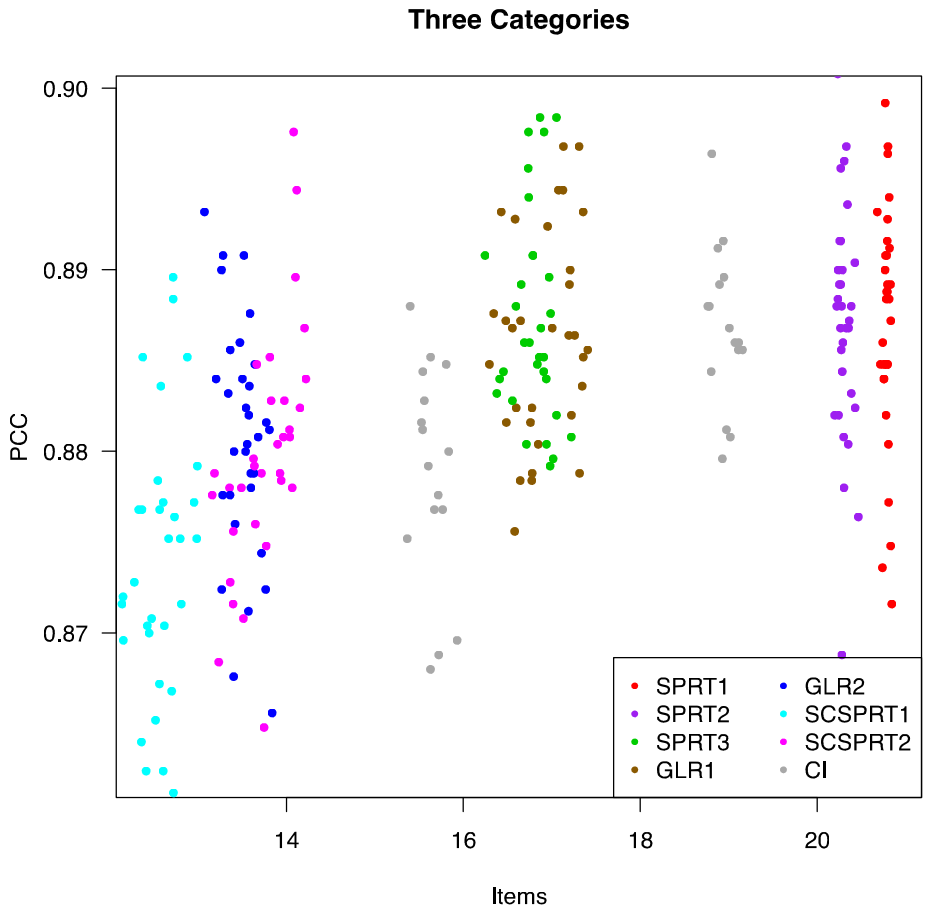

Figure 5

Proportion classified correctly and number of items on a three category classification task. Individual colors represent particular termination conditions, and points within a color indicate combinations of: item selection, exposure control, and ability estimation methods. SPRT1 and GLR1 uses $\alpha = \beta = .05$ and $\delta = .10$; SPRT 2 uses $\alpha = \beta = .10$ and $\delta = .10$; SPRT3 and GLR2 uses $\alpha = \beta = .10$ and $\delta = .20$; SCSPRT1 is without a conservative correction; SCSPRT2 uses a confidence interval correction of 67%; CI is based off a 95% coverage rate.

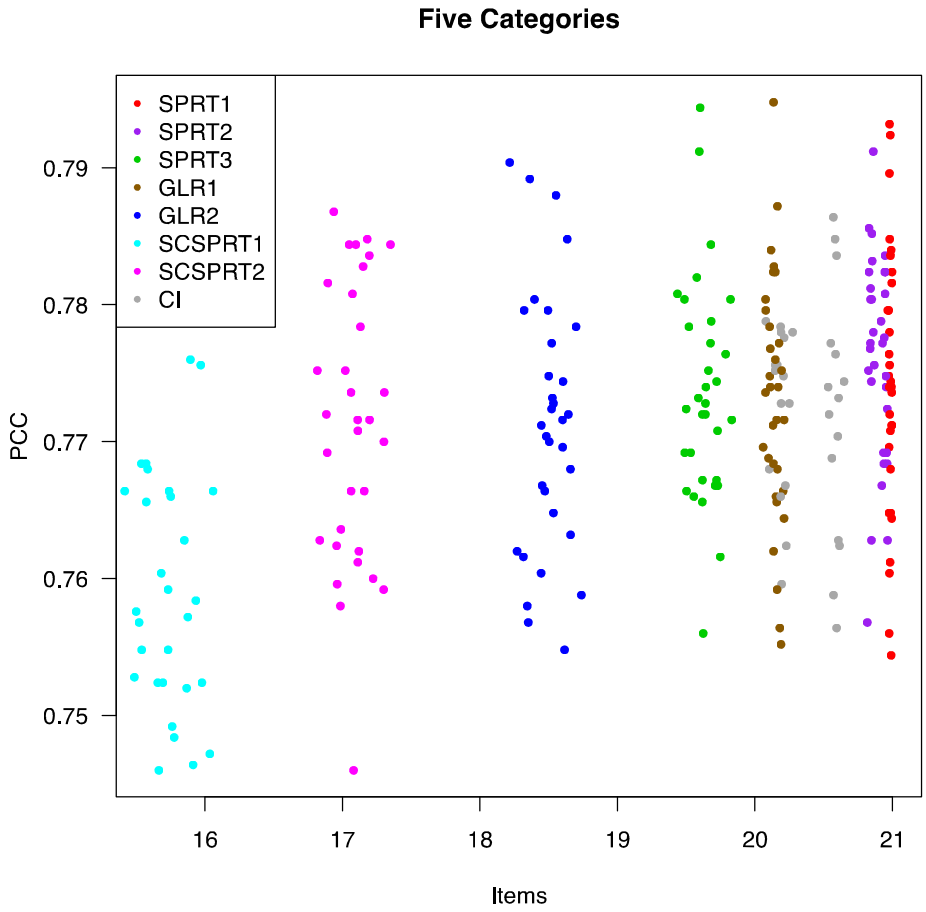

Figure 6

Proportion classified correctly and number of items on a five category classification task. Individual colors represent particular termination conditions, and points within a color indicate combinations of: item selection, exposure control, and ability estimation methods. SPRT1 and GLR1 uses $\alpha = \beta = .05$ and $\delta = .10$; SPRT 2 uses $\alpha = \beta = .10$ and $\delta = .10$; SPRT3 and GLR2 uses $\alpha = \beta = .10$ and $\delta = .20$; SCSPRT1 is without a conservative correction; SCSPRT2 uses a confidence interval correction of 67%; CI is based off a 95% coverage rate.

Table 1

Percent classified correctly and test length on a three category classification task. For test length, the 1st Quartile and mean number of items are displayed, whereas for percent classified correctly, the 3rd Quartile and mean classification rate are displayed. SPRT1 and GLR1 uses α = β = .05 and δ = .10; SPRT 2 uses α = β = .10 and δ = .10; SPRT3 and GLR2 uses α = β = .10 and δ = .20; SCSPRT1 is without a conservative correction; SCSPRT2 uses a conservative correction of .67; CI is based off a 95% coverage rate.

	SPRT1	*SPRT2*	*SPRT3*	*GLR1*	*GLR2*	*SCSPRT1*	*SCSPRT2*	*CI*
1Q Length	20.8	20.3	16.7	16.6	13.4	12.4	13.5	15.6
Mean Length	20.8	20.3	16.8	16.9	13.5	12.6	13.8	17.3
Mean PCC	.887	.887	.888	.886	.881	.874	.880	.883
3Q PCC	.891	.890	.891	.891	.884	.877	.882	.887

Table 2

Percent classified correctly and test length on a five category classification task. For test length, the 1st Quartile and mean number of items are displayed, whereas for percent classified correctly, the 3rd Quartile and mean classification rate are displayed. SPRT1 and GLR1 uses α = β = .05 and δ = .10; SPRT 2 uses α = β = .10 and δ = .10; SPRT3 and GLR2 uses α = β = .10 and δ = .20; SCSPRT1 is without a conservative correction; SCSPRT2 uses a conservative correction of .67; CI is based off a 95% coverage rate.

	SPRT1	*SPRT2*	*SPRT3*	*GLR1*	*GLR2*	*SCSPRT1*	*SCSPRT2*	*CI*
1Q Length	21.0	20.8	19.6	20.1	18.4	15.6	17.0	20.2
Mean Length	21.0	20.9	19.6	20.1	18.5	15.7	17.1	20.4
Mean PCC	.775	.776	.773	.773	.771	.758	.771	.773
3Q PCC	.782	.782	.779	.779	.778	.766	.781	.778

To reinforce conclusions drawn from Figures 5 and 6, we constructed four ANOVA tables: (1) Table 3 summarizes the effect of different factors on classification accuracy for a three category classification task; (2) Table 4 summarizes the effect of different factors on test length for a three category classification task: (3) Table 5 summarizes the effect of different factors on classification accuracy for a five category classification task; and (4) Table 6 summarizes the effect of different factors on test length for a five category classification task. We only examined main effects and those two-way interactions that include the stopping rule factor. Using an ANOVA is arguably inappropriate considering the dependent variable and research design, but we are only using it as a descriptive measure of variance accounted for by each factor (e.g., Guyer & Weiss, 2009).

Table 3

The sums of squares and η^2 = SSF/SST, where SSF is the sum of squares of a particular factor, for an ANOVA explaining mean classification accuracy for the three category, classification task. The ANOVA was run with all main effects and those interactions that relate to the termination factor.

Variance Type	**Sum of Squares**	η^2
Termination	0.00528	.330
Item Selection	0.00311	.195
Termination by Item Selection	0.00176	.110
Exposure	0.00037	.023
Termination by Ability Estimation	0.00028	.017
Termination by Exposure	0.00016	.010
Ability Estimation	0.00005	.003
Residuals	0.00498	.312
Total	**0.15971**	

Table 4

The sums of squares and η^2 = SSF/SST, where SSF is the sum of squares of a particular factor, for an ANOVA explaining mean test length for the three category, classification task. The ANOVA was run with all main effects and those interactions that relate to the termination factor.

Variance Type	**Sum of Squares**	η^2
Termination	2098.727	.9548
Termination by Item Selection	73.959	.0336
Item Selection	22.041	.0100
Exposure	1.502	.0007
Termination by Exposure	0.420	.0002
Termination by Ability Estimation	0.369	.0002
Ability Estimation	0.004	.0000
Residuals	1.057	.0005
Total	**2198.066**	

Based on Tables 3 through 6, termination, item selection, and the interaction of termination by item selection account for most of the variance of both percentage classified correctly and test length regardless of number of categories. Interestingly, termination condition appears to have both a smaller effect on classification accuracy and a larger effect on test length in the five category case (Tables 5 and 6) than in the three category case (Tables 3 and 4), suggesting that stochastic curtailment is advantageous for classification problems with multiple categories. Moreover, only in the three category task does item selection appear to contribute much to overall classification accuracy. Based on plots (not shown), a major reason for the moderately high η^2 for item selection is that selecting items at the classification bound results in slightly increased test length and classification accuracy. Therefore, more research should be undertaken to determine the nearest classification bound in models with guessing parameters. Termination accounts for most of the variance in test length (η^2 = .955 for the three category task and η^2 = .995 for the five category task), which is not surprising considering that we chose conditions already known to affect test length. Other than stopping rule and item

selection, all factors and interactions acount for so little variability in classification accuracy and test length that those conditions are not discussed.

Table 5

The sums of squares and η^2 = SSF/SST, where SSF is the sums of squares of a particular factor, for an ANOVA explaining mean classification accuracy for the five category, classification task. The ANOVA was run with all main effects and those interactions that relate to the termination factor.

Variance Type	**Sum of Squares**	η^2
Termination	0.00739	.263
Termination by Item Selection	0.00240	.085
Item Selection	0.00176	.063
Termination by Ability Estimation	0.00083	.029
Exposure	0.00081	.029
Termination by Exposure	0.00040	.014
Ability Estimation	0.00000	.000
Residuals	0.01453	.517
Total	**0.28116**	

Table 6

The sums of squares and η^2 = SSF/SST, where SSF is the sum of squares of a particular factor, for an ANOVA explaining mean test length for the five category, classification task. The ANOVA was run with all main effects and those interactions that relate to the termination factor.

Variance Type	**Sum of Squares**	η^2
Termination	817.714	.9951
Termination by Item Selection	1.732	.0021
Item Selection	0.625	.0008
Termination by Ability Estimation	0.549	.0007
Termination by Item Exposure	0.310	.0004
Exposure	0.274	.0003
Ability Estimation	0.116	.0001
Residuals	0.393	.0005
Total	**821.713**	

Tables 7 and 8 indicate the five best and five worst conditions in terms of item exposure. Termination conditions and ability estimation conditions affect item exposure mostly through differing test lengths. Not surprisingly, SCSPRT without a correction results in the best item exposure rates due to the shortest, average test length, and SPRT

(in general) results in the worst item exposure rates. In terms of item selection, the top five item exposure rates are associated with selecting items at the ability estimate, whereas the bottom five item exposure rates are associated with selecting items at the classification bounds. The reason for the stark disparity in item exposure rates when varying the location of selection is obvious: selecting items at the classification bounds overexposes those items with difficulty parameters close to the classification bounds. Unfortunately, we calibrated the Sympson-Hetter parameters identically for all of the conditions based on adaptively selecting items using maximum Fisher information due to time constraints. We suspect that re-calibrating the Sympson-Hetter parameters based on selecting items at the nearest cut-point would reduce the overexposure for items close to the classification bound. However, by using a modified set of Sympson-Hetter parameters using item selection at the nearest cut-point, the gains in item exposure control should be offset by increased test lengths and decreased classification accuracy. Moreover, we did not implement very strict item exposure controls. Using a maximum exposure rate of .1 rather than .2 should result in increased test length for all conditions.

Table 7

The lowest five maximum item exposure rates when classifying simulees into three and five categories, along with the specific overlap control, item selection, estimation, and termination conditions that led to those conditions. Also displayed are the calculated test overlap rates for each condition. The number in parentheses indicates how many maximum items were randomly chosen between to determine the next item on any CAT.

Lowest Item Overlap Conditions (Three Categories)

Overlap Control	Selection	Estimation	Termination	Max Exposure	Test Overlap
SH	Fisher at Theta (5)	MLE	SCSPRT 1	0.151	0.147
SH	KL at Theta (5)	MLE	SCSPRT 1	0.154	0.148
SH	KL at Theta (5)	WLE	SCSPRT 1	0.157	0.149
SH	Fisher at Theta (5)	WLE	SCSPRT 1	0.159	0.148
SH	Fisher at Theta (1)	MLE	SCSPRT 2	0.165	0.152

Lowest Item Overlap Conditions (Five Categories)

Overlap Control	Selection	Estimation	Termination	Max Exposure	Test Overlap
SH	Fisher at Theta (5)	MLE	SCSPRT 1	0.218	0.143
SH	Fisher at Theta (1)	MLE	SCSPRT 1	0.219	0.147
SH	KL at Theta (5)	MLE	SCSPRT 1	0.222	0.143
SH	KL at Theta (1)	WLE	SCSPRT 1	0.222	0.150
SH	KL at Theta (5)	WLE	SCSPRT 1	0.222	0.147

Table 8

The worst five maximum item exposure rates when classifying simulees into three and five categories, along with the specific overlap control, item selection, estimation, and termination conditions that led to those conditions. Also displayed are the calculated test overlap rates for each condition. The number in parentheses indicates how many maximum items were randomly chosen between to determine the next item on any CAT.

Highest Item Overlap Conditions (Three Categories)

Overlap Control	Selection	Estimation	Termination	Max Exposure	Test Overlap
None	Fisher at Bound (1)	WLE	SPRT 2	0.604	0.229
None	Fisher at Bound (1)	WLE	SPRT 1	0.600	0.232
None	KL at Bound (1)	MLE	SPRT 1	0.600	0.230
None	KL at Bound (1)	MLE	SPRT 3	0.600	0.196
None	KL at Bound (1)	WLE	SRPT 2	0.592	0.229

Highest Item Overlap Conditions (Five Categories)

Overlap Control	Selection	Estimation	Termination	Max Exposure	Test Overlap
None	KL at Bound (1)	WLE	Conf Int	0.505	0.151
None	KL at Bound (1)	WLE	SPRT 1	0.504	0.151
None	KL at Bound (1)	MLE	SPRT 3	0.500	0.152
None	KL at Bound (1)	MLE	GLR 1	0.500	0.150
None	KL at Bound (1)	WLE	SPRT 3	0.490	0.150

8. Discussion and Conclusions

Our study compared the classification accuracy and test length of various stopping rules using items culled from a real item bank and classifying examinees into either three or five categories. Stochastically curtailed SPRT improved over the truncated SPRT in terms of test length, and the accuracy trade off was small as long as we used a confidence interval correction. Moreover, in both tasks, the generalized likelihood ratio with a relatively large indifference region was nearly identical in test length and classification accuracy to stochastic curtailment with a 67% confidence interval correction. Because stochastic curtailment is computationally intensive, one should be able to use a GLR stopping rule without affecting accuracy and test length.

Even though the above simulations demonstrated improvement for the GLR and SCSPRT over the standard SPRT in many situations, there were a few limitations. For instance, we only varied a small set of parameters within each stopping rule. Difference between the SPRT and the GLR could be partly due to using the same indifference region in both procedures. For example, would an indifference region with half-width $\delta = .3$ for the SPRT result in similar performance to an indifference region with half-width $\delta = .2$ for the GLR? If the SPRT resulted in slightly higher classification accuracy with half-width $\delta = .3$, part of the accuracy decrement in the GLR could be due to using the same critical values as was used in the SPRT, and a more sophisticated simulation method (e.g., Bartroff et al., 2008) might be needed to determine appropriate thresholds for simulation.

Finkelman (2010) recently proposed variations on stochastic curtailment that relate more to generalized hypotheses. A simple alternative to stochastic curtailment is to

find the probability of the generalized likelihood ratio surpassing a critical threshold. None of the procedures proposed by Finkelman (2010) have been adapted to a classification task with multiple categories. Furthermore, most generalizations of multiple category sequential decision procedures are ad hoc implementations of Sobel and Wald (1949). It is clear that the critical values from the typical SPRT are inappropriate when there are many categories as evidenced by the classification rates in the mid .80s for the three category classification task and the mid .70s for the five category classification task. Several researchers have proposed individual critical values based either on the distance between categories (Spray, 1993) or a step-down procedure using a rank ordering of the likelihood ratio test statistic (e.g., Bartroff & Lai, 2010). Other researchers have extended sequential testing to multiple composite hypotheses (e.g., Pavlov, 1998), but these have yet to be applied to adaptive testing.

The simulations presented above demonstrate the power of likelihood ratio-based methods for efficiently and accurately classifying examinees when there are multiple categories. Adaptive testing is already being touted as the future of testing methodologies, and determining the most effective stopping rule is an important component of any CAT program. In light of both the accuracy and efficiency of adaptive testing procedures, the Common Core State Standards will soon adopt computerized adaptive tests in high-stakes exams (e.g., Way et al., 2010). Yet only when practitioners are knowledgeable of the ideal testing procedures across all assessment types will CAT fulfill its promise of being "highly compatible with the concept of vertically aligned standards and curricula that progress toward college and career readiness" (Way et al., 2010, p. 4).

References

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

Bartroff, J., & Lai, T. L. (2010). Multistage tests of composite hypotheses. *Communications in Statistics – Theory and Methods, 39*, 1597-1607.

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. 397–472). Reading, MA: Addison-Wesley.

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

Chang, Y. I. (2004). Application of sequential probability ratio test to computerized criterion-referenced testing. *Sequential Analysis, 23*, 45-61.

Chen, S.-Y., & Lei, P.-W. (2005). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement, 29*, 204-217.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 249-260.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*, 713-734.

Finkelman, M. (2003). *An adaptation of stochastic curtailment to truncate Wald's SPRT in computerized adaptive testing* (Tech. Rep.). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Finkelman, M. (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.

Guyer, R. D., & Weiss, D. J. (2009). Effect of early misfit in computerized adaptive testing on the recovery of theta. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved November 30, 2011 from: www.psych.umn.edu/psylabs/CATCentral

Kalohn, J. C., & Spray, J. C. (1999). The effect of model misspecification on classification decisions made using a computerized test. *Journal of Educational Measurement, 36*, 47-59.

Keener, R. W. (2010). *Theoretical Statistics: Topics for a Core Course*. New York, NY: Springer.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New Horizons in Testing: Latent Trait Theory and Computerized Adaptive Testing* (pp. 237-254). New York: Academic Press.

Kullback, S. (1959). *Information Theory and Statistics*. New York, NY: John Wiley and Sons.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79-86.

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

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14,* 367-386.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48,* 233-245.

Pavlov, I. V. (1988). A sequential procedure for testing many composite hypotheses. *Theory of*

Probability and its Applications, 33, 138-142.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-255). New York, NY: Academic Press.

Sobel, M., & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Annals of Mathematical Statistics, 20*, 502-522.

Spray, J. A. (1993). *Multiple-category classification using a sequential probability ratio test* (Tech. Rep.). Iowa City: ACT Research Report Series.

Spray, J. A., & Reckase, M. D. (1994). *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.

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

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item-exposure rates in computerized adaptive testing*. Paper presented at the annual meeting of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation, 12*, 1-13.

Thompson, N. A. (2009a). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*, 778-793.

Thompson, N. A. (2009b). Using the generalized likelihood ratio as a termination criterion. In

D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved June 29, 2011 from www.psych.umn.edu/psylabs/CATCentral

Thompson, N. A. (2010, June). *Nominal error rates in computerized classification testing*. Paper presented at the first annual conference of the International Association for Computerized Adaptive Testing, Arnhem Netherlands.

Wainer, H. (2000). *Computerized Adaptive Testing: A Primer*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

Way, W. D., Twing, J. S., Camera, W., Sweeney, K., Lazar, S., & Mazzeo, J. (2010, February). *Some considerations relating to the use of adaptive testing for the Common Core Assessments*. Retrieved November 14, 2011, from the College Board Web site: http://professionals.collegeboard.com/profdownload/some-considerations-use-of-adaptive-testing.pdf

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

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

Welch, R. E., & Frick, T. W. (1993). Computerized adaptive testing in instructional settings. *Educational Technology Research and Development, 41*, 47-62.

Wouda, J. T., & Eggen, T. J. H. M. (2009). Computerized classification testing in more than two categories by using stochastic curtailment. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved June 7, 2011 from www.psych.umn.edu/psylabs/CATCentral/

Yang, X., Poggio, J.C., & Glasnapp, D.R. (2006). Effects of estimation bias on multiple category

classification with an IRT-based adaptive classification procedure. Effects of Estimation Bias on *Educational and Psychological Measurement, 66,* 545-564.