Accuracy and Efficiency in Classifying Examinees using Computerized Adaptive Tests

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Steven Warren Nydick

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF ARTS

Niels Waller

Thesis Advisor

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 enough or when the maximum number of items has been reached. In short tests, the SPRT is inefficient due to properties of the log-likelihood ratio test statistic, necessitating other methods that address shortcomings of the SPRT, such as 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. Most of the work on stopping rules was derived for the special case of a CCT with only two categories. 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. A follow up study is proposed to develop and compare Bayesian decision methods in classification CAT. Because Bayesian methods use posterior probabilities, they should better control the overall error rate when there are more than two categories and the prior distribution of examinee ability is known.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the enactment of No Child Left Behind (NCLB; 2008) and the imminent implementation of the Common Core State Standards (CCSS; 2010), test developers must take pains to assure the reputability of every test score. Considering the quantity of examinees, the high cost of assessment, and the consequence of misclassification, test implementers require procedures that are both accurate and efficient. Tests that continuously adapt to the estimated ability level of each examinee increase accuracy by individually determining when enough information has been collected to end each test, and they reduce inefficiency by preventing unnecessarily easy or difficult items from being wasted on inappropriate examinees. In light of both accuracy and efficiency, the CCSS will soon adopt computerized adaptive tests (CAT; e.g., Wainer, 2000; Weiss, 1982) in high stakes exams (e.g., Way, Twing, Camara, Sweeney, Lazar, & Mazzeo, 2010).

Although a primary goal of NCLB is to track changes in proficiency at the individual level, many summative assessments must also classify examinees into pre-specified categories. Classification tests are used to assess everything from job qualifications (e.g., ACT, 2007), teacher certification (e.g., Pearson, 2011), and student mastery. The most basic classification task is to compare an examinee's score to the minimal score needed to pass an exam (e.g., Kingsbury & Weiss, 1983; Welch & Frick, 1993; Yang, Poggio, & Glasnapp, 2006) and is usually referred to as a "mastery" or "certification" test (Bejar, 1983). However, many classification tests attempt to place examinees into one of a number of categories, such as coarsely defined ability or job skill level, with consequences for career advancement, success, and earning potential. Computerized classification testing (CCT) is a subset of computerized adaptive testing with the intent of assigning examinees to mutually exclusive categories. Unlike CATs designed for equiprecise measurement (i.e., tests in which each test score is measured with equal precision; Weiss, 1982), the procedures implemented in CCT aim only to increase the accuracy and efficiency of classification. That is to say, the optimal CCT algorithm selects items (e.g., Eggen, 1999) and terminates a test (e.g., Eggen & Straetmans, 2000; Finkelman, 2008a; Lewis & Sheehan, 1990) based on category membership rather than individual ability. Because CCT requires decision rules to choose between competing categories, psychometricians have applied sequential methods taken from statistical decision theory (Wald, 1947) to adaptive classification algorithms (e.g., Eggen, 1999; Finkelman, 2003, 2008a; 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 (i.i.d.) data have been collected to choose between one of two simple hypotheses (Keener, 2010, p. 417–422). With regard to classification testing, these simple hypotheses are quantified as specific ability levels within each category (e.g., Reckase, 1983; Spray, 1993; Spray & Reckase, 1996). With two categories, the SPRT must continuously decide whether there is enough evidence to classify the examinee in the lower category, whether there is enough evidence to classify the examinee in the upper category, or whether the examinee should be administered another item. As will be described in Section 2.2, evidence for classification in the SPRT depends on the relationship between test score and classification hypotheses.

Primary justification for using the SPRT in sequentially planned experiments is its optimality with i.i.d. data, as presented in the Wald-Wolfowitz theorem: Given two simple hypotheses, H_0 and H_1, "of all tests with the same power the sequential probability ratio test requires on the average fewest observations" (Wald & Wolfowitz, 1948, p. 326). Unfortunately, even assuming that item responses for each examinee are independent, items on an exam are of various difficulty, so that the identically distributed assumption required for the optimality of the SPRT does not hold. Reckase (1979) noticed that "the model as presented assumes that the probability of a correct response is the same for all items in the pool" (p. 84), but his only concern in violating identically distributed data was as an impediment to providing closed form solutions for the expected test length. Unlike Reckase, Finkelman (2008a, 2008b) illustrated that classical SPRT is not optimal as a termination criterion in the case of unbalanced item discriminations and/or a ceiling on the possible number of items in a test. To circumvent limitations with the SPRT, Finkelman (2003, 2008a, 2010) provided alternative termination criteria designed to take advantage of unique CAT properties. Specifically, he added variations on a probabilistic stopping rule taken from clinical trials known as stochastic curtailment (Lan, Simon, & Halpern, 1982). The Sequential Probability Ratio Test with Stochastic Curtailment (SCSPRT) makes a classification decision based, in part, on information from the remaining items in the bank. Other modifications of the original SPRT procedure include: (1) Using composite hypotheses that take into consideration an examinee's current ability estimate (Generalized Likelihood Ratio or GLR; Bartroff, Finkelman, & Lai, 2008; Thompson, 2009, 2010); and (2) Extending the SPRT, SCSPRT, and GLR in CCT to more than two categories (e.g., Spray, 1993; Eggen, 1999; Wouda & Eggen, 2009).

To date, researchers have evaluated the relative performance of the SPRT with only single alternatives (either SCSPRT or GLR) in situations requiring examinees to be classified in two or three categories. The purpose of this study was to provide practicable equations for the SCSPRT under the three-parameter logistic IRT model and to compare the SPRT, GLR, and SCSPRT in a classification task with more than three categories. The remainder of this thesis is organized as follows. In Chapter 2, I identify and review the three-parameter logistic IRT model underlying many computerized classification tests, explain the SPRT as a long-standing solution to classification decisions, and introduce alternative criteria (i.e., the SCSPRT and GLR) that ostensibly circumvent limitations of the SPRT in CCT. In Chapter 3, I outline a simulation study designed to compare each of the above termination criteria in realistic testing situations. In Chapter 4, I summarize results from the simulation, and in Chapter 5, I evaluate the overall simulation, discuss limitations, and propose future directions.

Chapter 2

Background on Classification Testing

2.1 Item Response Theory and Classification Testing

Item response theory (IRT) is often used to formalize the relationship between responses to individual test items and examinee ability. Outside of the context of specialty applications (e.g., DiBello & Stout, 2007 for cognitive diagnosis models; Goegebeur, De Boeck, Wollack, & Cohen, 2008 for models based on speeded tests; Revuelta, 2008 for models representing item solving processes), the most popular IRT model remains the unidimensional, binary, three-parameter logistic model (3PL; Birnbaum, 1968) or simplifications thereof. Specifically, let θ denote the latent variable underlying responses to test items (we will refer to this variable as "ability" even though any unidimensional latent variable would be appropriate), assume that item responses are conditionally independent[1] given a particular level of θ (call that value θ_i for examinee i), and allow

[1] I prefer the phrase "conditionally independent" to "locally independent" due to dependencies among items selected for each computerized adaptive test (e.g., Mislevy & Chang, 2000).

all items to have two possible choices, only one of which is correct. Then, according to the 3PL, the probability of person i responding to item j in the keyed direction can be represented with the following item response function (IRF):

$$p_j(\theta_i) = \Pr(U_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + \exp[-Da_j(\theta_i - b_j)]} \tag{2.1}$$

where θ_i denotes "ability" for person i, b_j determines the inflection point of the IRF, a_j is proportional to the slope of the IRF at its inflection point, c_j indicates the lower asymptote (i.e., the probability of a person with infinitely low "ability" correctly answering item j), and D is a scaling constant usually specified to be either 1 or 1.702[2]. In IRT parlance, b_j is referred to as the difficulty or extremity parameter, a_j as the discrimination parameter, and c_j as the pseudo-guessing parameter. Furthermore, because D is a scaling constant that does not affect model fit, it is subsequently absorbed into the discrimination parameter for clarity.

If a single ability underlies responses to test items, then classification categories can be defined as ranges of values on this latent dimension. For instance, consider a test with one a priori cut-point, θ_0, separating masters from non-masters. True classification depends on the location of an examinee's ability in relation to the cut-point. If θ_i (the true, latent ability of examinee i) is greater than θ_0, then the examinee is deemed a master, and any other decision represents a "false negative" or Type II error (Finkelman, 2008a). If, on the other hand, θ_i is less than θ_0, then the examinee is a non-master and any other classification represents a "false positive" or Type I error. Of course, examinees can only take a limited number of items, so every decision is made with incomplete information, and any method that can terminate a test early must ensure

[2] Originally, the probability of response was defined using a normal-ogive item response function, and $D = 1.702$ is the estimator that minimizes that maximum difference between the logistic IRF and the normal-ogive IRF with the same item parameter values (Camilli, 1994)

that the decision would be unlikely to change after administering more items. Sequential tests were derived to use the fewest items while strictly controlling the classification error rate, so I shall presently describe each of the sequential criteria commonly used in CCT.

2.2 Classification Criteria in CCT

Most modern termination criteria in computerized classification testing are modifications of Wald's Sequential Probability Ratio Test (SPRT; Wald, 1947). Thererfore, I will first explain the SPRT as applied to CAT and then present the GLR and SC-SPRT as alternative termination criteria that were designed to circumvent many of the shortcomings in the original SPRT.

2.2.1 The Sequential Probability Ratio Test

The classic SPRT in classification CAT (e.g., Eggen, 2000; Reckase, 1983; Spray & Reckase, 1996) starts by defining a simplified form of the classification problem. Given a choice between two categories separated by a cut-point, θ_0, hypotheses are specified as the ends of an indifference region surrounding θ_0. To illustrate this idea, let the indifference region be symmetric, and denote the half-width of the indifference region as δ. Then point hypotheses can be specified as

$$\begin{aligned} \mathrm{H}_0 &: \theta_i = \theta_0 - \delta \\ \mathrm{H}_1 &: \theta_i = \theta_0 + \delta \end{aligned}$$

where H_0 is a surrogate for all θ_i below $\theta_0 - \delta$, and H_1 is a surrogate for all θ_i above $\theta_0 + \delta$. Furthermore, any true "ability" within the indifference region, such that $\theta_i \in (\theta_0 - \delta, \theta_0 + \delta)$, can be classified in either category.

After an examinee responds to an item on the classification test, the SPRT determines whether the null hypothesis should be accepted, the alternative hypothesis should be accepted, or the examinee should be administered another item. To make a decision, the SPRT compares the likelihood ratio test statistic to critical values based on pre-specified error rates. For instance, if responses are conditionally independent and follow a unidimensional, binary, item response function, then the log-likelihood for a single examinee given a particular response pattern, $\mathbf{u}_i = [u_{i1}, u_{i2}, \ldots, u_{iJ}]^T$, is

$$\log[L(\theta|\mathbf{u}_i)] = \sum_{j=1}^{J} \Big[u_{ij} \log[p_j(\theta)] + (1 - u_{ij}) \log[1 - p_j(\theta)]\Big] \tag{2.2}$$

with $p_j(\theta)$ defined in Equation (2.1) and $L(\theta|\mathbf{u}_i)$ denoting the likelihood function for θ given response vector $\mathbf{u}_i$. Adopting two point hypotheses of examinee "ability", $\mathrm{H}_0 : \theta_l = \theta_0 - \delta$ and $\mathrm{H}_1 : \theta_u = \theta_0 + \delta$, the log-likelihood ratio of an examinee evidencing ability θ_u relative to θ_l is

$$\log\Big[LR(\theta_u, \theta_l|\mathbf{u}_i)\Big] = \log\left[\frac{L(\theta_u|\mathbf{u}_i)}{L(\theta_l|\mathbf{u}_i)}\right] = \log\Big[L(\theta_u|\mathbf{u}_i)\Big] - \log\Big[L(\theta_l|\mathbf{u}_i)\Big]. \tag{2.3}$$

When Equation (2.3) is a large, positive number, there is sizable evidence for θ_u (rather than θ_l) as generating the particular response pattern, $\mathbf{u}_i$. Conversely, when Equation (2.3) is a large, negative number, there is sizable evidence supporting θ_l.

Justification for using a likelihood ratio test statistic when testing simple hypotheses is based on the Neyman-Pearson lemma (Casella & Berger, 1990, p. 366). According to Neyman-Pearson, for a fixed sample size, N, and conditional on a particular Type I error rate, α, the uniformly most powerful (UMP) test rejects H_0 contingent on the size of the likelihood ratio test statistic. The Wald-Wolfowitz theorem generalizes Neyman-Pearson to the case of optional stopping. Specifically, let

$X_1, X_2, \ldots$ be an i.i.d. sample from common density f with unknown parameter(s) $\boldsymbol{\theta}$. Then given simple hypotheses, $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ versus $\mathrm{H}_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_2$, and critical values, A and B, where $0 < A < B < \infty$, a rule that stops sampling when $N = \inf\left\{n \geq 1 \ : \ \prod_{i=1}^{n}\left[\frac{f(x_i|\boldsymbol{\theta}_1)}{f(x_i|\boldsymbol{\theta}_2)}\right] \leq A \quad \text{or} \quad \prod_{i=1}^{n}\left[\frac{f(x_i|\boldsymbol{\theta}_1)}{f(x_i|\boldsymbol{\theta}_2)}\right] \geq B\right\}$ is optimal in the set of all tests with the same Type I and Type II error rates (Lai, 1997). In the case of sequential tests, the critical values, A and B, are usually selected to yield specific α (Type I error rate) and β (Type II error rate) levels. Even though methods exist for determining A and B given desired error rates, Wald (1947) recommended using $A = \frac{\beta}{1-\alpha}$ and $B = \frac{1-\beta}{\alpha}$ as easily calculable, although approximate, bounds[3].

Using the theory of sequential tests as a template, psychometricians have developed a simple procedure for terminating classification CATs:

1. Pick α and β, and define $C_l = \log(A) = \log[\beta/(1-\alpha)]$ and $C_u = \log(B) = \log[(1-\beta)/\alpha]$. Determine the minimum number of items, j_{min}, the maximum number of items, j_{max}, and set the current item to $j_{\text{tmp}} = 1$.

2. Administer item j_{tmp}. If $j_{\text{tmp}} < j_{\text{min}}$, set $j_{\text{tmp}} = j_{\text{tmp}} + 1$ and repeat step (2). Otherwise, proceed to step (3).

3. Calculate $C = \log\Big[LR(\theta_u, \theta_l|\mathbf{u}_i)\Big]$ as defined in Equation (2.3). If $C < C_l$, accept H_0 (i.e., classify the examinee in the lower category) and terminate the test. If $C > C_u$, accept H_1 (i.e., classify the examinee in the upper category) and terminate the test. If $C \in [C_l, C_u]$ and $j_{\text{tmp}} < j_{\text{max}}$, set $j_{\text{tmp}} = j_{\text{tmp}} + 1$, return to step (2). If $C \in [C_l, C_u]$ and $j_{\text{tmp}} = j_{\text{max}}$, proceed to step (4).

4. Accept H_0 if $C \leq \frac{C_l + C_u}{2}$ or accept H_1 if $C > \frac{C_l + C_u}{2}$, and terminate the test.

[3] Wald (1945) showed that $\alpha' \leq \frac{\alpha}{1-\beta}$ and $\beta' \leq \frac{\beta}{1-\alpha}$, where α' and β' are the actual type I and type II error rates using $A = \frac{\beta}{1-\alpha}$ and $B = \frac{1-\beta}{\alpha}$. Because $\alpha' + \beta' \leq \alpha + \beta$, at most one error rate will exceed its desired value.

Using the above methodology, individually tailored classification can be designed with controlled misclassification rates. Note that when the true value of θ_i is outside of the indifference region, the actual Type I and Type II error rates are smaller than the nominal error rates when $\theta_i = \theta_u$ or $\theta_i = \theta_l$ (e.g., Chang, 2004, p. 49).

Although standard SPRT is useful for mastery tests, many tasks need to classify examinees into more than two categories. To accomodate these tasks, Sobel and Wald (1949) generalized the SPRT to choose between 1 of 3 parameter values for μ when the underlying data are normally distributed, which they conjectured was "not far from being optimum" (p. 503). For classification CAT with K categories, Sobel and Wald's method can be extended as follows.

Calculate the log-likelihood ratio values: $C_0, C_1, \ldots, C_{K-1}, C_K$, where K indicates the number of point values tested, C_k is the log-likelihood ratio of $\mathrm{H}_0 : \theta = \theta_k$ versus $\mathrm{H}_1 : \theta = \theta_{k+1}$, θ_k is the kth smallest proposed value of θ, $C_K = -\infty$, and $C_0 = \infty$. Then another observation should be collected if there exists some $k \in \{1, \ldots, K\}$ such that $C_k \in [C_l, C_u]$, and data collection should cease and θ_k chosen as the estimated parameter value if for some $k \in \{1, \ldots, K\}$, $C_{k-1} > C_u$ and $C_k < C_l$. Essentially, the proposed extension of Sobel and Wald classifies an examinee into a category if his/her trait estimate is determined to be above the lower bound and below the upper bound of that category. Sobel and Wald (1949) asserted that when C_l and C_u are fixed and there are only three categories, then it is impossible for two hypotheses to be accepted on the same sequential step. Not surprisingly, this natural extension of the SPRT has been readily applied to classification tests (e.g., Eggen, 1999; Eggen & Straetsmans, 2000; Spray, 1993; Wouda & Eggen, 2009).

Unfortunately, researchers have identified at least two limitations of the classic SPRT in adaptive testing. First, although Wald and Wolfowitz (1948) proved optimality of the SPRT when testing simple hypotheses, practitioners are seldom interested only in the endpoints of an indifference region. Classification tasks are usually used to determine whether examinees evince one of an infinite number of trait levels between any two cut-points. Thus, it is essential that termination criteria assess a range of "ability." In light of this concern, the Generalized Likelihood Ratio (GLR; Bartroff, Finkelman, & Lai, 2008; Thompson, 2009, 2010) was proposed as a simple modification of the SPRT that tests composite hypotheses. Second, although the SPRT controls the error rate for infinitely long experiments, every realistic CAT must end. After $j_{\max}$ items, the CCT terminates regardless of whether or not $C_k \notin [C_l, C_u]$ for every category, and classifies the examinee into the category corresponding to the highest log-likelihood ratio (Eggen, 1999). Finkelman (2003, 2008a) demonstrated that the SPRT with Truncation (TSPRT) is inefficient given fixed levels of α and β, and he proposed a modified procedure that estimates the probability of examinee i switching categories by $j_{\max}$. In the next two subsections, I explore this idea further and elaborate the required adjustments to the SPRT algorithm.

2.2.2 The Generalized Likelihood Ratio Test

The Generalized Likelihood Ratio (GLR) is a modification of the original SPRT algorithm for testing composite hypotheses. Consider a hypothesis about real-valued parameter $\theta \in \mathbb{R}^1$ from density $f(x|\theta)$, such that $\mathrm{H}_0 : \theta \leq \theta_0$ and $\mathrm{H}_1 : \theta > \theta_0$. Then, in many situations, using a generalized likelihood ratio test statistic with $L(\theta_1|\mathbf{x}) = \arg\max_{\theta > \theta_0}\{f(\mathbf{x}|\theta)\}$ in the numerator and $L(\theta_2|\mathbf{x}) = \arg\max_{\theta \leq \theta_0}\{f(\mathbf{x}|\theta)\}$ in the denominator results in a uniformly most powerful (UMP) test (Casella & Berger, 1990, p. 368). Replacing either θ_1 or θ_2 with the maximum likelihood estimate of θ is a natural extension of the simple likelihood ratio procedure for testing composite hypotheses. Intuitively, the modified approach compares $\hat{\theta}_{\mathrm{MLE}}$ to the most likely value of the composite hypothesis to which $\hat{\theta}_{\mathrm{MLE}}$ does not belong. For notational convenience, let $\hat{\theta} = \hat{\theta}_{\mathrm{MLE}}$

when the meaning of the term is obvious.

Due to purportedly optimal properties, generalized likelihood ratio statistics have been adopted in sequential hypotheses testing (e.g., Lai, 1997, 2001; Lai & Shih, 2004). As Lai (2001), wrote, “simulation studies and asymptotic analysis have shown that [the number of items needed to make a decision using a GLR] is nearly optimal over a broad range of parameter values θ, performing almost as well as [a procedure] that assumes θ to be known” (p. 307). Because of desirable characteristics, Bartroff, Finkelman, and Lai (2008) proposed appropriating the test statistic

$$\log\left[GLR(\theta_0|\mathbf{u}_i)\right] = \log\left[L(\hat{\theta}|\mathbf{u}_i)\right] - \log\left[L(\theta'|\mathbf{u}_i)\right] \tag{2.4}$$

as an alternative to the likelihood ratio in CCT, where $\mathrm{H}_0 : \theta \leq \theta_0$, $\mathrm{H}_1 : \theta > \theta_0$, and θ' has not yet been defined but depends on the value of $\hat{\theta}$. The initial Generalized Likelihood Ratio Test (GLR; Bartroff et al., 2008) for terminating adaptive tests was based off of the Haybittle-Peto procedure described by Lai and Shih (2004). The gist of the GLR can be described as follows. At each iteration of the exam, estimate θ by $\hat{\theta}$. If $\hat{\theta} > \theta_0$, calculate $G = \log\left[GLR(\theta_0|\mathbf{u}_i)\right]$ using Equation (2.4) by setting $\theta' = \theta_0 - \delta$. If $\hat{\theta} \leq \theta_0$, calculate G by setting $\theta' = \theta_+^{j_{\max}}$ [4]. Finally, compare G to $G_l^{(j_{\max})}$ and $G_u^{(j_{\max})}$, where $G_l^{(j_{\max})}$ and $G_u^{(j_{\max})}$ are critical values determined prior to the study. Because Equation (2.4) will always be positive, choose H_0 if $G > G_l^{(j_{\max})}$ and $\hat{\theta} \leq \theta_0$, choose H_1 if $G > G_u^{(j_{\max})}$ and $\hat{\theta} > \theta_0$, or continue testing if the relevant critical value is not exceeded. If $j = j_{\max}$, terminate the test, and classify the examinee into the category in which $\hat{\theta}$ falls. Bartroff et al. (2008) described a method of determining $G_l^{(j_{\max})}$, $G_u^{(j_{\max})}$, and $\theta_+^{(j_{\max})}$ based on the maximum number of items, $j_{\max}$, and the desired error rates, α and β.

[4] $\theta_+^{j_{\max}}$ is chosen so that the Type I and Type II error rates are as specified at $j_{\max}$. Bartroff et al. (2008) called $\theta_+^{j_{\max}}$ the “implied alternative” (p. 476) and estimated it via simulation.

Because procedures for determining exact $\theta_+^{(j_{\max})}$ and critical values for the GLR are complicated, Thompson (2009, 2010) suggested that the GLR be identical to "the fixed point SPRT, with the exception that θ_1 and θ_2 [in the generalized likelihood ratio test statistic] are allowed to vary" (Thompson, 2010, p. 5). Rather than calculating Equation (2.4) and comparing it to $G_l^{(j_{\max})}$ and $G_u^{(j_{\max})}$, Thompson advised calculating

$$\log\Big[GLR(\theta_l, \theta_u|\mathbf{u}_i)\Big] = \sup_{\theta_1 \geq \theta_l}\bigg(\log\Big[L(\theta_1|\mathbf{u}_i)\Big]\bigg) - \sup_{\theta_2 \leq \theta_u}\bigg(\log\Big[L(\theta_2|\mathbf{u}_i)\Big]\bigg) \tag{2.5}$$

after item $j \in \{j_{\min}, \ldots, j_{\max}\}$ and comparing the result to $C_l = \log[\beta/(1-\alpha)]$ and $C_u = \log[(1-\beta)/\alpha]$ as in the SPRT. Even when using an inexact system, he found that the classification accuracy was not noticeably different than the Truncated SPRT, but the average test length was reduced by approximately 10 items on a $j_{\max} = 200$ item CCT (Thompson, 2010). Yet the original SPRT and the modified GLR are, at best, nearly optimal when $j_{\max} = \infty$. Finkelman (2003, 2008a) showed that for $j_{\max} < \infty$, a procedure that also terminated a CCT when the probability of switching categories before $j = j_{\max}$ was small would statistically dominate the SPRT given the same α, β, and δ. Finkelman based this supplementary stopping rule, which is called stochastic curtailment, on the work of Lan, Simon, and Halpern (1982) from the clinical trials literature. In the next subsection, I describe the logic behind stochastic curtailment as applied to classification CAT.

2.2.3 The SPRT with Stochastic Curtailment

The curtailed version of a sequential procedure (Eisenberg & Ghosh, 1980) makes decision D_k before $j = j_{\max}$ (where $j_{\max}$ is the maximum number of observations) if and only if for every $s \neq k$, decision D_s can never be chosen. Eisenberg & Ghosh (1980) proved that for a UMP test with fixed sample size, the curtailed version must also be UMP. At a minimum, a curtailed sequential procedure performs at least as well as its truncated cousin. Because the criterion for curtailment is usually difficult to obtain, researchers have modified the curtailed stopping rule to make decision D_k before $j = j_{\max}$ if and only if for every $s \neq k$, the probability of choosing decision D_s is below some threshold (Finkelman, 2008a, p. 453). As applied to CCT, the Sequential Probability Ratio Test with Stochastic Curtailment (SCSPRT; Finkelman, 2003) classifies an examinee when the probability of switching categories by maximum test length is small[5]. That is to say, after item $j_{\max}$, the TSPRT will force a decision and classify examinee i into category k, where $\hat{\theta}_i \in [\theta_{k-1}, \theta_k]$. However, even though the likelihood ratio test statistic might not satisfy the SPRT criterion by item $j_{\text{tmp}} < j_{\max}$, as described in Subsection 2.2.1, a different classification might be unlikely by item $j_{\max}$ due to a weak set of remaining items. The SCSPRT proceeds in three steps: (1) SPRT, (2) Curtailment, and (3) Stochastic Curtailment.

Curtailment

In sequential testing, curtailment means prematurely ending data collection when outstanding observations cannot change the ultimate decision. The curtailment problem can be redefined in terms of the log-likelihood of the item response function, as expressed in Equation (2.2). The maximum likelihood estimate for examinee i, $\hat{\theta}_i$, is determined by setting the derivative of the log-likelihood equal to 0 and solving for θ. After j_{tmp} items, the derivative of the log-likelihood function for any unidimensional, binary response model can be written as

[5] Finkelman (2003) and Finkelman (2008a) used slightly different criteria for SCSPRT. Finkelman (2003) classified examinees when the probability of switching categories was small, whereas Finkelman (2008a) classified examinees when the probability of the log-likelihood ratio surpassing a predefined, maximum test length critical value, C_m was small. I chose the former for the purposes of this paper, as one would only set $C_m \neq 0$ if certain classification errors were more important.

$$\frac{d\log[L(\theta|\mathbf{u}_i^{(j_{\text{tmp}})})]}{d\theta} = \sum_{j=1}^{j_{\text{tmp}}} \left\{ [u_{ij} - p_j(\theta)] \frac{p'_j(\theta)}{p_j(\theta)[1 - p_j(\theta)]} \right\} \tag{2.6}$$

where

$$p'_j(\theta) = (1 - c_j)a_j p^1_j(\theta)[1 - p^1_j(\theta)] \tag{2.7}$$

is the derivative of the 3PL item response function (Equation 2.1) and

$$p^1_j(\theta) = \frac{1}{1 + \exp[-a_j(\theta - b_j)]} \tag{2.8}$$

is the probability of responding in the keyed direction using a similar, two-parameter logistic model (2PL). Setting Equation (2.6) equal to 0 and replacing p'_j with Equations (2.7) and (2.8) results in

$$\sum_{j=1}^{j_{\text{tmp}}} a_j p^1_j(\theta) = \sum_{j=1}^{j_{\text{tmp}}} \left[a_j u_{ij} \left(\frac{p^1_j(\theta)}{p_j(\theta)} \right) \right]. \tag{2.9}$$

Several parts of Equation (2.9) merit comment. First, the maximum likelihood estimate is such that that right-hand-side of Equation (2.9) equals the left-hand-side. However, the expected value of both sides with respect to some θ will always be equal. Second, when using a model without guessing parameters, Equation (2.9) simplifies to

$$\sum_{j=1}^{j_{\text{tmp}}} a_j p_j(\theta) = \sum_{j=1}^{j_{\text{tmp}}} a_j u_{ij}$$

so that $\sum_{j=1}^{j_{\text{tmp}}} a_j u_{ij}$ is a sufficient statistic for θ_i. Finally, $\left(\frac{p^1_j(\theta)}{p_j(\theta)}\right)$ on the far right of Equation (2.9) can be rewritten as

$$p_j^{c_j}(\theta) = \frac{p_j^1(\theta)}{p_j(\theta)} = \frac{\exp[a_j(\theta - b_j)]}{c_j + \exp[a_j(\theta - b_j)]} \tag{2.10}$$

where c_j denotes the lower asymptote for item j. Equation (2.10), a generalized logistic curve, is an increasing function of θ for $c_j > 0$ with a minimum of 0 and a maximum of 1.

After each item in an adaptive test, the maximum likelihood estimate of θ is found by solving Equation (2.9) for θ. Whether a classification test should be curtailed depends on whether particular values of θ can ever be maximum likelihood estimates. For instance, let $\hat{\theta}_i^{(j_{\mathrm{tmp}})} < \theta_0$, where θ_0 is the bound separating the lower and the upper categories and $\hat{\theta}_i^{(j_{\mathrm{tmp}})}$ is the maximum likelihood estimate of θ_i after j_{tmp} items. Then a test should terminate at $j = j_{\mathrm{tmp}}$ if Equation (2.9) cannot hold for $\theta > \theta_0$ at $j_{\max}$ items. An equivalent statement of curtailment is that if the derivative of the log-likelihood function can not equal 0 for all values of ability in the upper category by $j = j_{\max}$, then the remaining items provide no relevant information for classifying examinee i.

Equation (2.9) suggests an approximate method of curtailment. Define

$$T_{j_{\max}} = \sum_{j=1}^{j_{\max}} a_j p_j(\theta_0) \tag{2.11}$$

where $T_{j_{\max}}$ is the left-hand-side of Equation (2.9) for $\theta = \theta_0$ after $j = j_{\max}$ items. $T_{j_{\max}}$ is recalculated after each item $j = j_{\mathrm{tmp}}$ by combining items $j = 1, \ldots, j_{\mathrm{tmp}}$ from the current CAT with hypothetical items $j = j_{\mathrm{tmp+1}}, \ldots, j_{\max}$ from the remaining items in the bank. Next define

$$S_{j_{\mathrm{tmp}}} = \sum_{j=1}^{j_{\mathrm{tmp}}} u_{ij} a_j p_j^{c_j}(\tilde{\theta}_i) \tag{2.12}$$

where $\tilde{\theta}_i$ is the closest endpoint of a $100\gamma\%$ confidence interval usually based on a normal approximation (described in Subsection 2.2.3). Then a classification test should terminate if

$$T_{j_{\max}} > S_{j_{\text{tmp}}} + \sum_{j=j_{\text{tmp}}+1}^{j_{\max}} a_j p_j^{c_j}(\tilde{\theta}_i) \tag{2.13}$$

or if

$$T_{j_{\max}} < S_{j_{\text{tmp}}}. \tag{2.14}$$

When Equation (2.13) holds, examinee i can answer the remaining exam items correctly without $\hat{\theta}_i^{(j_{\max})} > \theta_0$, and when Equation (2.14) holds, examinee i can answer the remaining exam items incorrectly without $\hat{\theta}_i^{(j_{\max})} < \theta_0$. In both cases, additional information will not change the classification. Even though strict curtailment is guaranteed to dominate the truncated SPRT, a probabilistic procedure might work better in practice.

Stochastic Curtailment

The curtailment procedure, outlined in the preceding subsection, requires the probability of the examinee changing categories by item $j = j_{\max}$ to be nearly 0 before terminating the classification test. Finkelman (2003) modified the curtailment condition so that the test will terminate if the probability of category reassignment by item $j = j_{\max}$ is small. Assume that $j_{\text{remain}} = j_{\max} - j_{\text{tmp}}$ is large, so that

$$S_{j_{\max}}|\mathbf{u}_i^{(j_{\text{tmp}})} = S_{j_{\text{tmp}}} + \sum_{j=j_{\text{tmp}}+1}^{j_{\max}} u_{ij} a_j p_j^{c_j}(\tilde{\theta}_i) \tag{2.15}$$

is approximately normally distributed. Based on the curtailment procedure, the probability that the classification decision will not change is approximately equivalent to the probability that $S_{j_{\max}}|\mathbf{u}_i^{(j_{\text{tmp}})} - T_{j_{\max}}$ does not cross 0. Finkelman (2003) demonstrated that when $\hat{\theta}_i^{(j_{\text{tmp}})} > \theta_0$, this approximation can be expressed as

$$P(\hat{\theta}_i^{(j_{\max})} > \theta_0|\hat{\theta}_i^{(j_{\text{tmp}})}) \approx \Phi\left(\frac{\hat{E}(S_{j_{\max}}|\mathbf{u}_i^{(j_{\text{tmp}})}) - T_{j_{\max}}}{\widehat{\text{Var}}(S_{j_{\max}}|\mathbf{u}_i^{(j_{\text{tmp}})})}\right) \tag{2.16}$$

and when $\hat{\theta}_i^{(\text{tmp})} \leq \theta_0$, then

$$P(\hat{\theta}_i^{(j_{\max})} \leq \theta_0|\hat{\theta}_i^{(j_{\text{tmp}})}) \approx \Phi\left(\frac{T_{j_{\max}} - \hat{E}(S_{j_{\max}}|\mathbf{u}_i^{(j_{\text{tmp}})})}{\widehat{\text{Var}}(S_{\max}|\mathbf{u}_i^{(j_{\text{tmp}})})}\right) \tag{2.17}$$

where $\Phi(q) = \Pr(Z \leq q)$ and $Z \sim N(0,1)$. It can easily be shown that

$$\hat{E}(S_{j_{\max}}|\mathbf{u}_i^{(j_{\text{tmp}})}) = S_{j_{\text{tmp}}} + \sum_{j=j_{\text{tmp}}+1}^{j_{\max}} a_j p_j^1(\tilde{\theta}_i) \tag{2.18}$$

and

$$\widehat{\text{Var}}(S_{j_{\max}}|\mathbf{u}_i^{(j_{\max})}) = \sum_{j=j_{\text{tmp}}+1}^{j_{\max}} [a_j p_j^1(\tilde{\theta}_i)]^2 \left(\frac{1 - p_j(\tilde{\theta}_i)}{p_j(\tilde{\theta}_i)}\right) \tag{2.19}$$

based off of Equations (2.10) and (2.15). Note that $\tilde{\theta}_i$ is the endpoint of a $100\gamma\%$ confidence interval closest to θ_0 (described in Subsection 2.2.3). Finkelman (2003) recommended using an Edgeworth expansion for small j_{remain}. Edgeworth expansions better approximate the distribution of a finite sum than the normal distribution (e.g., Blinnikov & Moessner, 1998; Weng, 2010), although, to my knowledge, nobody has applied this improvement for estimating probabilities in classification problems.

In the original paradigm, Finkelman (2003, 2008a) proposed that after administering any item past $j_{\min}$, the algorithm should do the following (given a nominal error rate of some $\epsilon < .50$):

1. Check the SPRT criterion (as described in Subsection 2.2.1) and whether $j_{\text{tmp}} =$

$j_{\max}$. If either criterion is satisfied, end the test and classify the examinee. Otherwise, go on to step (2).

2. Check the curtailment condition (as described in Subsection 2.2.3). If the examinee will not change categories by maximum test length, end the test and classify the examinee. Otherwise, go on to step (3).

3. Calculate either Equation (2.16) (if $\hat{\theta}_i^{(j_{\text{tmp}})} > \theta_0$) or Equation (2.17) (if $\hat{\theta}_i^{(j_{\text{tmp}})} \leq \theta_0$). If the cumulative density is greater than $1 - \epsilon$, terminate the test and classify the examinee into the category in which $\hat{\theta}_i^{(j_{\text{tmp}})}$ is located. Otherwise, administer another item.

Finkelman (2003) was not clear on how to adapt the SCSPRT beyond a mastery test with a known item order, and many classification tasks have more than two categories. In the next subsection, I describe an adaptation of stochastic curtailment when there are multiple categories.

SCSPRT with Many Categories

In Subsection 2.2.1, I outlined a logical adjustment of the SPRT procedure when there are more than two categories, based on by Sobel and Wald (1949): Check the SPRT criterion at every bound, and classify examinee i into category k if there is evidence that he or she is above category $k - 1$ and below category $k + 1$. Unlike the SPRT, Wouda and Eggen (2009) noted that the "[SCSPRT] stopping rules need information of items that could be administered in the future ... in order to be able to assign examinees to a certain trait level" (p. 6). Both $T_{j_{\max}}$ (Equation 2.11) and $\hat{E}(S_{j_{\max}} | \mathbf{u}_i^{(j_{\text{tmp}})})$ (Equation 2.18) depend not *merely* on the j_{tmp} items administered prior to calculating them, but also on the remaining $j_{\text{remain}} = j_{\max} - j_{\text{tmp}}$ items from an examinee's hypothetical test. To approximate the remaining items, Finkelman (2008a, 2010) recommended using a "representative set [of items] expected to be similar to the items ultimately chosen" (Finkelman, 2010, p. 33) and described several procedures (e.g., maximizing Fisher Information or Kullback-Leibler divergence at the classification bound; Chang & Ying, 1996) that completely identify the set of future items. A reasonable solution (e.g., Wouda & Eggen, 2009) is to calculate Equations (2.11) and (2.18) using items that maximize Fisher information around particular locations.

The Fisher Information function (FI; Lord, 1980) for item j can be written

$$\begin{aligned}\mathcal{I}_j(\theta) = -E\left[\frac{\partial^2 \log[L(\theta|\mathbf{u})]}{\partial\theta^2}\right] &= \frac{[p_j'(\theta)]^2}{p_j(\theta)[1-p_j(\theta)]} \\ &= \frac{a_j^2(1-c_j)}{\Big(c_j+\exp[a_j(\theta-b_j)]\Big)\Big(c_j+\exp[-a_j(\theta-b_j)]\Big)^2}\end{aligned} \tag{2.20}$$

where $p_j(\theta)$ is defined in Equation (2.1) and $p_j'(\theta)$ is defined in Equation (2.7). Fisher information measures the curvature of the log-likelihood function in a small area surrounding the maximum likelihood estimate and relates to the asymptotic variance of $\hat{\theta}$ (e.g., Frank, 2009, who also compared FI to Shannon Entropy in the context of evolutionary biology). Choosing items to maximize Fisher information at particular points is a classic (e.g., Weiss, 1982), easily calculated, and hardly outdone (e.g., Chen, Ankenmann, & Chang, 2000) item selection algorithm in adaptive testing. When there is one cut-point, θ_0, separating two categories, the SCSPRT algorithm would:

1. Estimate $T_{j_{\max}}$ by picking the top j_{remain} items that maximize Equation (2.20) at θ_0.

2. Estimate $\hat{E}(S_{j_{\max}}|\mathbf{u}_i^{(j_{\text{tmp}})})$ by picking the top j_{remain} items that maximize Equation (2.20) based on the confidence interval endpoint closest to θ_0.

When there are multiple cut-points and $\hat{\theta}_i^{(j_{\text{tmp}})}$ is between cut-points θ_{k-1} and θ_k, the

SCSPRT algorithm would need to find two probabilities: (1) The probability that $\hat{\theta}_i$ will exceed θ_k by item $j_{\max}$; and (2) The probability that $\hat{\theta}_i$ will fall below θ_{k-1} by item $j_{\max}$. Only if *both* probabilities are below a threshold should the SCSPRT algorithm terminate with θ_i classified in category k. To calculate the probability that $\hat{\theta}_i^{(j_{\max})}$ will exceed θ_k, the SCSPRT algorithm would estimate $T_{j_{\max}}$ by picking the top j_{remain} items that maximize Equation (2.20) at θ_k and $\hat{E}(S_{j_{\max}}|\mathbf{u}_i^{(j_{\text{tmp}})})$ by picking the top j_{remain} items that maximize Equation (2.20) at

$$\tilde{\theta}_{i,u}^{(j_{\text{tmp}})} = \hat{\theta}_i^{(j_{\text{tmp}})} + \frac{z_{(1+\gamma)/2}}{\sqrt{\sum_{j=1}^{j_{\text{tmp}}} \mathcal{I}_j(\hat{\theta}_i^{(j_{\text{tmp}})})}} \tag{2.21}$$

where γ is the desired confidence level and $z_{(1+\gamma)/2}$ is the $\frac{1+\gamma}{2}$ quantile of a standard normal distribution. To calculate the probability that $\hat{\theta}_i^{(j_{\max})}$ will fall below θ_{k-1}, the SCSPRT algorithm would estimate $T_{j_{\max}}$ by picking the top j_{remain} items that maximize Equation (2.20) at θ_{k-1} and $\hat{E}(S_{j_{\max}}|\mathbf{u}_i^{(j_{\text{tmp}})})$ by picking the top j_{remain} items that maximize Equation (2.20) at

$$\tilde{\theta}_{i,l}^{(j_{\text{tmp}})} = \hat{\theta}_i^{(j_{\text{tmp}})} - \frac{z_{(1+\gamma)/2}}{\sqrt{\sum_{j=1}^{j_{\text{tmp}}} \mathcal{I}_j(\hat{\theta}_i^{(j_{\text{tmp}})})}}. \tag{2.22}$$

Note that both $\tilde{\theta}_{i,l}^{(j_{\text{tmp}})}$ and $\tilde{\theta}_{i,u}^{(j_{\text{tmp}})}$ depend only on items already administered to examinee i. Unlike Wouda and Eggen (2009), $\tilde{\theta}_i$ has been represented as one end of a two-sided confidence interval due to multiple bounds being tested (and $\tilde{\theta}_i$ needing to take on multiple values) within the same sequential step. Using the procedures just described, I am now prepared to outline a simulation study designed to assess the relative performance of each termination criterion in classifying examinees into more than three categories.

Chapter 3

Current Study Design

In this chapter, I describe a simulation study that was designed to compare classification rates, test lengths, and item exposure rates for the SPRT, GLR, and SCSPRT under a variety of conditions when there are more than three classification categories. I first discuss properties of the assessment instrument and latent trait distribution, and I then outline the ability estimation methods, item selection algorithms, exposure control conditions, and specific stopping rules that were used in the current study.

3.1 Assessment Properties

3.1.1 Assessment Instrument

A real item pool was employed, consisting of $J = 600$ operational items calibrated in Bilog MG-3 (Zimowski, Muraki, Mislevy, & Bock, 2006) under the assumption of a three-parameter logistic model with $D = 1.702$. Due to security concerns, the source of the item pool can not be revealed, and all of the a and b parameters were linearly transformed from their original metric. The transformed a-parameters had a mean of 1.20 and a standard deviation of 0.33; the transformed b-parameters had a mean of

0.06 and a standard deviation of 1.43; and the untransformed c-parameters had a mean of .15 and a standard deviation of .07. The four classification boundaries (separating five categories) were chosen to be similar to those used in live administration given the transformed a and b parameters: $\theta_1 = -1.39$, $\theta_2 = -0.47$, $\theta_3 = 0.28$, and $\theta_4 = 1.18$.

3.1.2 Latent Trait Distribution

Two simulations were performed. First, the conditional accuracy and test length was determined by simulating 400 classification tests at each of 81 θs evenly spaced between -4 and 4. Second, the aggregate classification accuracy and test length was estimated by simulating $N = 5000$ θs from a standard normal distribution. The distribution of θ was chosen to be similar to the empirical distribution for this particular assessment after a location and scale transformation.

3.1.3 Testing Procedure

Regardless of condition, there were a few restrictions on how items were to be administered to each simulee. Simulees were allowed to take between $j_{\min} = 8$ and $j_{\max} = 21$ items. Most of the previous studies (e.g., Batroff, et al., 2008; Spray & Reckase, 1996; Wouda & Eggen, 2009) used a maximum of $j_{\max} = 50$ items; however, the current simulation was designed as a pilot study for future use in live administration and constrained by the framework of the prospective CAT. Items were assigned to one of 8 content categories, every classification test must include at least one item from each category, and sequentially administered items must be from two different categories. Due to the sensitive nature of the assessment, both the number and distribution of items within each category is suppressed even though each could have affected overall classification.

3.2 Ability Estimation

Although equiprecise measurement was not the goal of this study, several item selection and classification algorithms require provisional estimates of examinee ability. Ability was estimated via one of two criteria: (1) The maximum likelihood estimate (MLE) or (2) The weighted likelihood estimate (WLE). Maximum likelihood estimation was described in Subsection 2.2.3.

Weighted likelihood estimation was proposed as a bias correction to the maximum likelihood estimate (e.g., Lord, 1983; Warm, 1989) and is found by solving for the root of a modified score function

$$\log[L(\theta|\mathbf{u}_i^{(j_{\text{tmp}})})]'_{\text{WLE}} = \sum_{j=1}^{j_{\text{tmp}}} \frac{[u_{ij} - p_j(\theta)]p'_j(\theta)}{p_j(\theta)[1 - p_j(\theta)]} + \frac{\sum_{j=1}^{j_{\text{tmp}}} \mathcal{H}_j(\theta)}{2\sum_{j=1}^{j_{\text{tmp}}} \mathcal{I}_j(\theta)} \tag{3.1}$$

where $p_j(\theta)$ is defined in Equation (2.1), and $p'_j(\theta)$ is defined in Equation (2.7). The correction factor, $\frac{\sum_{j=1}^{j_{\text{tmp}}} \mathcal{H}_j(\theta)}{2\sum_{j=1}^{j_{\text{tmp}}} \mathcal{I}_j(\theta)}$, depends on the Fisher information of the preceding items ($\mathcal{I}_j(\theta)$, Equation 2.20) and

$$\mathcal{H}_j(\theta) = \frac{p'_j(\theta)p''_j(\theta)}{p_j(\theta)[1 - p_j(\theta)]} \tag{3.2}$$

where $p''_j(\theta)$ is the second derivative of the 3PL item response function with respect to θ and can be written as

$$p''_j(\theta) = [p'_j(\theta)]' = a_j p'_j(\theta)[1 - p_j(\theta)]\Big[1 - \exp[a_j(\theta - b_j)]\Big]. \tag{3.3}$$

Although many classification studies have used $\hat{\theta}_{\text{MLE}}$ to estimate ability (e.g., Bartroff et al., 2008; Finkelman, 2008a), weighted likelihood estimation has gained a foothold in the classification literature (e.g., Eggen & Straetmans, 2000; Wouda & Eggen, 2009). Moreover, SCSPRT depends on precise estimation of ability to calculate the necessary probabilities, so it is conceivable that WLE could improve the classification accuracy of stochastic curtailment, especially early in a test.

3.3 Item Selection

To identify the effect of stopping rules on classification accuracy, two methods of item selection were applied: (1) maximum Fisher information at the current ability estimate; and (2) maximum Fisher information at the nearest cut-point. The nearest cut-point was determined by minimizing the critical inequality

$$Q = |S_{j_{\text{tmp}},2} - \delta(L_k + U_k)| \tag{3.4}$$

where δ is the half-width of the indifference region,

$$L_k = \frac{\log\left(\frac{\beta}{1-\alpha}\right) - \sum_{j=1}^{j_{\text{tmp}}} \log\left[\frac{1-p_j(\theta_k+\delta)}{1-p_j(\theta_k-\delta)}\right]}{2\delta}, \tag{3.5}$$

$$U_k = \frac{\log\left(\frac{1-\beta}{\alpha}\right) - \sum_{j=1}^{j_{\text{tmp}}} \log\left[\frac{1-p_j(\theta_k+\delta)}{1-p_j(\theta_k-\delta)}\right]}{2\delta}, \tag{3.6}$$

θ_k is the trait level corresponding to cut-point k, and

$$S_{j_{\text{tmp}},2} = \sum_{j=1}^{j_{\text{tmp}}} u_{ij} \left[\frac{c_j + \exp[a_j(\theta_k + \delta - b_j)]}{c_j + \exp[a_j(\theta_k - \delta - b_j)]}\right]. \tag{3.7}$$

Equations (3.5), (3.6), and (3.7) are rearrangements of the simple log-likelihood from the SPRT set to the lower (L_k) and the upper (U_k) critical values. Eggen (1999) admitted that selecting items at θ_k that minimizes Equation (3.4) was "an ad hoc criterion that is not based on theory" and "the generalization to the [3PL] ... is not as straightforward" (p. 258). However, he found that minimizing Equation (3.4) performed at least as well as other item selection criteria.

Regardless of condition, simulees were administered the first four items at random provided they satisfied minimal content constraints. Simulees who did not have a mixed response pattern after four items were randomly administered low difficulty or high difficulty items until a bounded, maximum likelihood estimate could be calculated, after which they were administered items according to maximum Fisher information.

3.4 Exposure Control

Due to security concerns, it was required to limit the proportion of examinees who responded to any given item. Controlling item exposure was implemented through the Sympson-Hetter (S-H; 1985) method. S-H is described in detail elsewhere (e.g., Chen & Lei, 2005), so only a brief explanation will be provided. Let S_j denote whether item j is selected for a random examinee, and let A_j denote whether item j is administered to that examinee. It is well known that

$$\Pr(A_j) = \Pr(A_j \cap S_j) = \Pr(A_j|S_j)\Pr(S_j), \tag{3.8}$$

where $\Pr(A_j) = \Pr(A_j \cap S_j)$ because items must be selected before they are administered. In most adaptive tests, selecting an item implies administering that item, so $\Pr(A_j) = \Pr(A_j|S_j)\Pr(S_j) = 1 \times \Pr(S_j) = \Pr(S_j)$. Sympson-Hetter controls the item exposure rate by reducing the proportion of selected items actually administered to examinees. Specifically, let $r_{\max}$ be the maximum percentage of examinees that should see any item. Then it is easy to simulate adaptive tests under appropriate testing conditions, track the proportion of times that each item is selected, and set

$$\Pr(A_j|S_j) = \min\left\{\frac{r_{\max}}{\Pr(S_j)},\ 1\right\}. \tag{3.9}$$

The S-H calibration was repeated 15 times to stabilize the estimate of $\Pr(A_j|S_j)$ and used as its simulees 5000 θs simulated from a standard normal distribution. Even though the first simulation used a uniform distribution of θ, the distribution of θ implemented in S-H was chosen to be standard normal to match the realistic distribution of examinees taking the CAT.

During CAT administration, each item is associated with a conditional probability of being administered should it be selected, $\Pr(A_j|S_j)$. After an item is selected, it is compared to a realization, u, of a random variable, U, where $U \sim \text{Unif}(0, 1)$. If $\Pr(A_j|S_j) > u$, the item is administered, but if $\Pr(A_j|S_j) \leq u$, then the item is not administered and subsequently removed from the list of available items. Three values of r_{max} were chosen: .1, .2, and 1, where $\text{r}_{\text{max}} = 1$ connotes no item exposure.

3.5 Stopping Rules

The ultimate goal was to compare the classification accuracy and test length of each stopping method under a variety of conditions. For all three methods, the nominal Type I and Type II error rates were set to $\alpha = \beta = .10$. Thompson (2010) found that "nominal PCC [i.e., percent classified correctly] had very little effect on observed PCC" (p. 9), and even though most researchers have used $\alpha = \beta = .05$ as nominal error rates (e.g., Bartroff, 2008; Finkelman, 2008a; Thompson, 2009), setting α and β to .05 might prevent the SPRT from typically classifying simulees before a maximum test length of $j_{\text{max}} = 21$ items. When classifying simulees using the SPRT or GLR, the half-width of the indifference region, δ, was varied between .10 to .30 in .05 increments. Previous researchers have set δ between .10 and .30, and Thompson (2010) determined that classification accuracy declined at about $\delta = .30$. When classifying simulees using the SCSPRT, δ was set to either .20 or .25 due to preliminary results suggesting that

$\delta > .25$ led to unacceptable classification accuracy, and $\hat{\theta}_{\text{MLE}}$ or $\hat{\theta}_{\text{WLE}}$ was corrected by $100\gamma = 0$, 38, 68, 86 and 95 percent confidence intervals. The chosen confidence levels roughly correspond to a $\tilde{\theta}$ that is 0, .5, 1.0, 1.5, and 2 standard errors closer to the classification bound than $\hat{\theta}_{\text{MLE}}$ or $\hat{\theta}_{\text{WLE}}$.

Chapter 4

Simulation Results

The results are summarized in two parts. I first describe the conditional test length and classification accuracy corresponding to various stopping rules, and I then summarize the overall test length and classification accuracy by assuming that latent ability is generated from a standard normal distribution. Both techniques address different research questions: (1) How well the stopping rules classify simulees near each cut-point; and (2) How well the stopping rules will perform given a particular set of examinees.

4.1 Results 1: Conditional on Specific Ability Values

The first set of simulations examined the accuracy and test length of various stopping rules conditional on particular ability values. Most stopping rules do not proclaim to classify all examinees with equal precision. Examinees who have ability near a cutting point should require more items to determine the appropriate classification category. To assess whether the termination criteria result in short and reasonably precise decisions, 400 CCT were simulated at each of 81 ability points evenly spaced between -4 and 4 for each combination of conditions. Due to time constraints, ability was estimated using

MLE and the stochastic curtailment conditions always used a half-width of $\delta = .20$.

Figure 4.1 presents the conditional test length for certain termination conditions when items were selected using maximum Fisher information at $\hat{\theta}_i$ and there were no item exposure constraints. Consider the dark and light blue curves, which represent SPRT with a large indifference region and GLR with a smaller indifference region. Although SPRT with a large indifference region ($\delta = .30$) tends to result in fewer items near the classification bounds than GLR with a smaller indifference region ($\delta = .20$), the improvement in test length does not extend to extreme values of θ_i. Moreover, notice that all of the termination conditions result in quicker classification when ability is in the extremes. However, when using 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 due to properties of the log-likelihood ratio test statistic in the three-parameter logistic model along with the items selected in any CAT. For the SPRT to be efficient, items need to be selected based, in part, on the location of the classification bound. To demonstrate the gain in efficiency, Figure 4.2 presents the conditional test length for the same termination conditions when items were selected based on maximum Fisher information at the nearest cut-point. The only noticeable difference from Figure 4.1 to Figure 4.2 is that average test length is more symmetrical around the lowest and highest cut-points when classifying examinees using the SPRT. Additional properties of the SPRT in the three-parameter logistic model are explored in Section 4.3.

Comparing test length of the SCSPRT conditions and the GLR conditions is not as straightforward as comparing the SPRT conditions. Based on Figures 4.1 and 4.2, stochastic curtailment without a confidence interval correction results in the shortest average tests for most levels of ability. Although average test length does not depend much on δ for extreme values of θ_i when using GLR as a stopping rule, the improvement

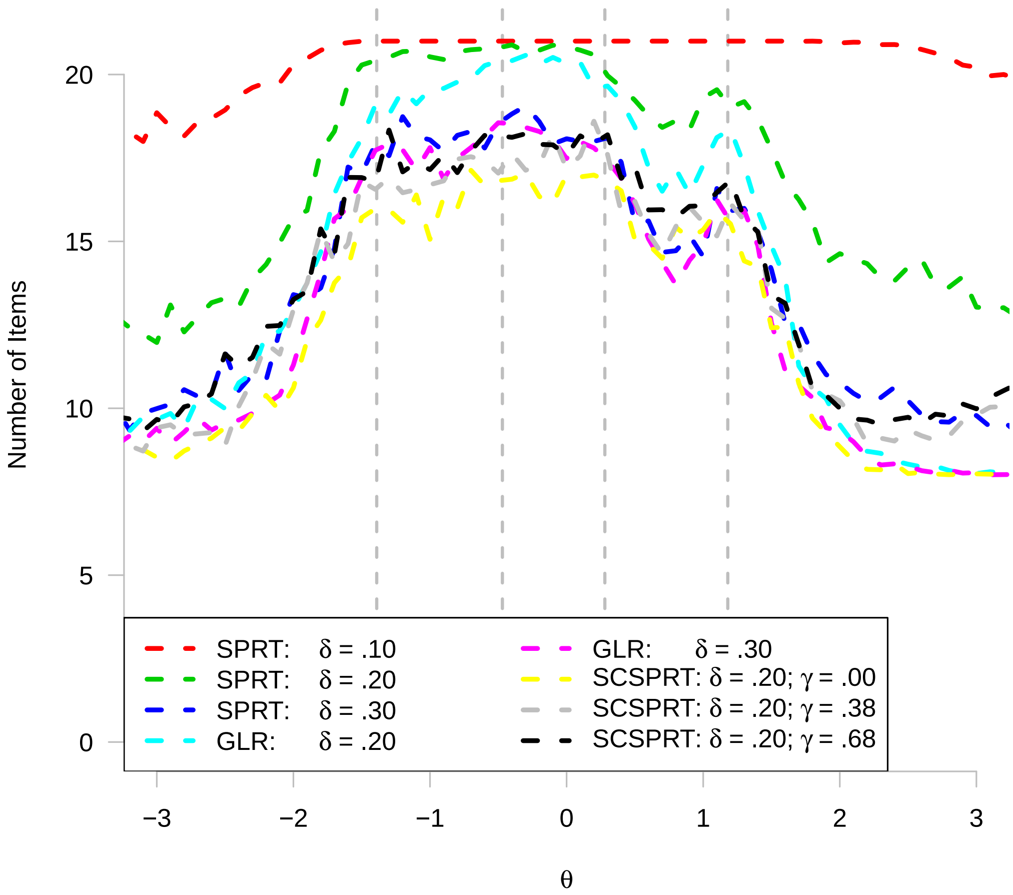


Figure 4.1: Test length averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at $\hat{\theta}_i$, ability estimated by maximum likelihood estimation, and no item exposure control. The vertical bars represent the classification bounds. Only a few termination conditions are presented for illustration purposes.

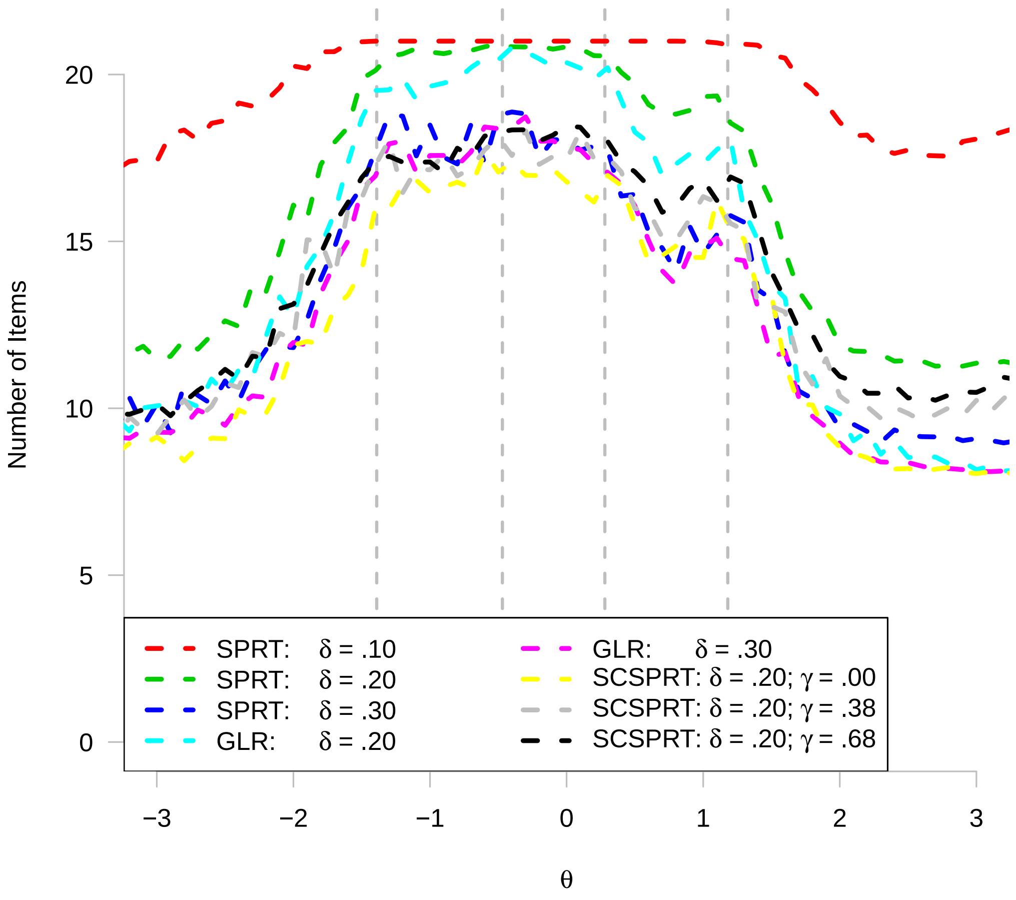


Figure 4.2: Test length averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at the nearest cut-point, ability estimated by maximum likelihood estimation, and no item exposure control. The vertical bars represent the classification bounds. Only a few termination conditions are presented for illustration purposes.

in test length when $\delta = .30$ as compared with $\delta = .20$ is noticeable for examinees near the classification bounds. A final thing to note is that when terminating classification tests based on stochastic curtailment and using a confidence interval correction, the average test length slopes upward at the extremes. When examinees have extreme values of θ_i, they 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.

Figures 4.3 and 4.4 display the conditional accuracy corresponding to the conditions discussed in Figures 4.1 and 4.2, and the form of both plots look nearly identical. All of the stopping rules appear to classify examinees with equal precision when true ability is above the highest classification bound, and stochastic curtailment appears to consistently perform worse than the other methods when examinee ability is below the highest classification bound. However, by using a probabilistic stopping rule, stochastic curtailment was expected to result in decreased classification accuracy. Importantly, the differences between the curves of Figures 4.3 and 4.4 are generally small, which supports the finding of Finkelman (2008) that stochastic curtailment does not much affect classification accuracy for large γ. As can also be seen, the biggest divergences in the accuracy curves are midway between the middle classification bounds. Typical simulations assume that θ follows a standard normal distribution, so under common conditions, the misclassification rate might be exacerbated for certain stopping rules due to most simulees being between the middle classification bounds. I discuss overall classification in the next subsection. Unfortunately, the specified Type I and Type II error rates only apply to θ_i close to (e.g., δ away from) each classification bound, and the differences between the curves at specific values of θ_i are difficult to discern. Those differences are magnified by constructing tables composed of test length and

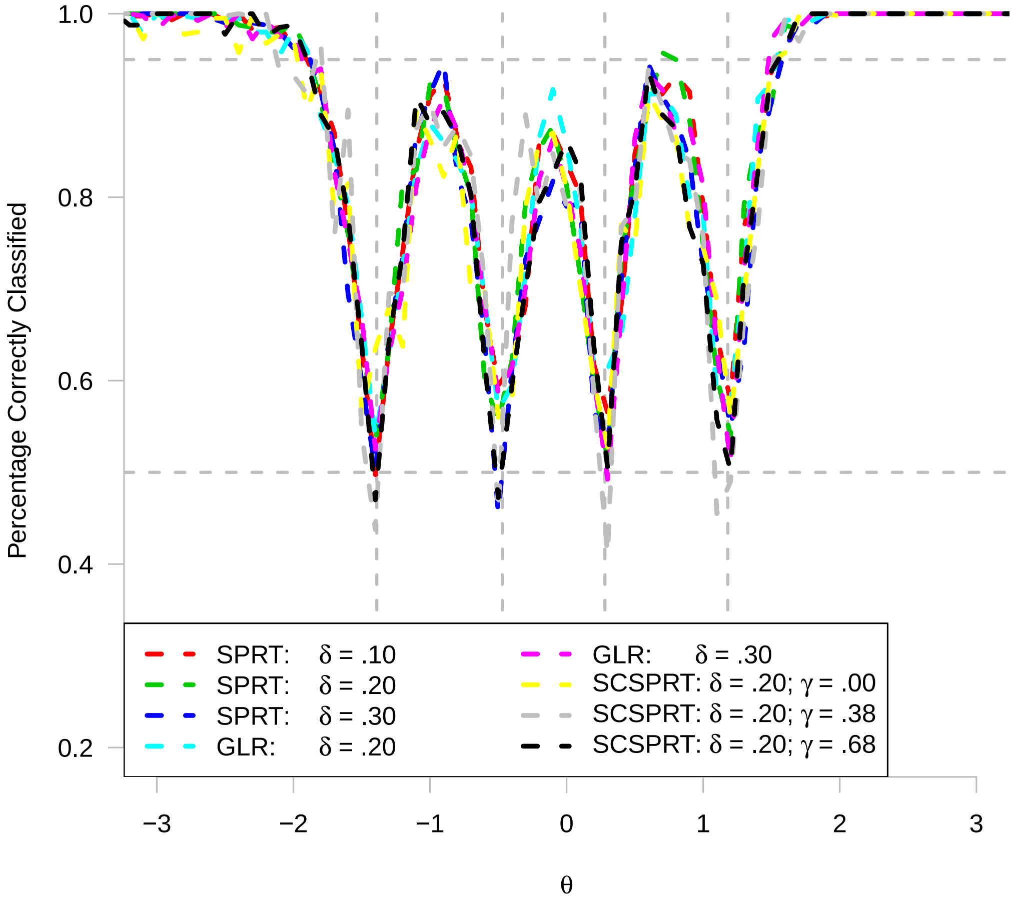


Figure 4.3: Classification accuracy averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at $\hat{\theta}_i$, ability estimated by maximum likelihood estimation, and no item exposure control. The vertical bars represent the classification bounds, and the horizontal bars 50% classification accuracy and 95% classification accuracy. Only a few termination conditions are presented for illustration purposes.

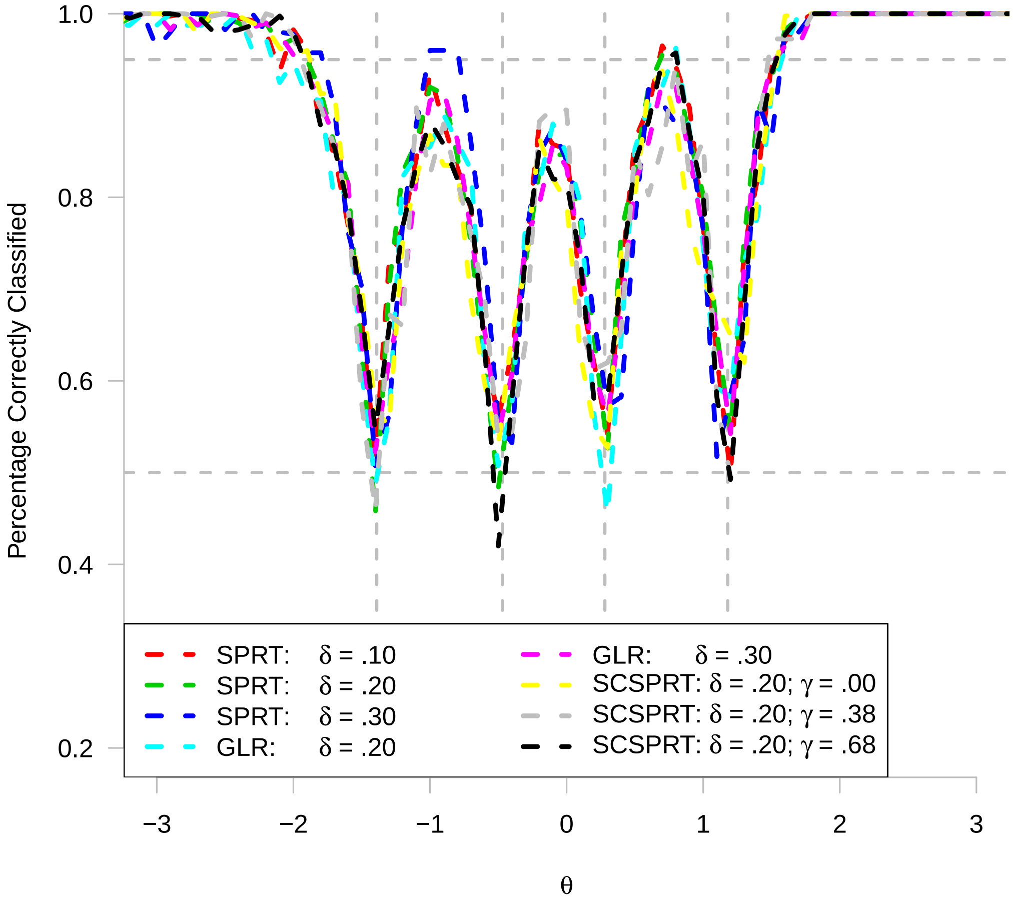


Figure 4.4: Classification accuracy averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at the nearest cut-point, ability estimated by maximum likelihood estimation, and no item exposure control. The vertical bars represent the classification bounds, and the horizontal bars 50% classification accuracy and 95% classification accuracy. Only a few termination conditions are presented for illustration purposes.

classification accuracy at specific true ability values.

Tables 4.1 and 4.2 indicate the average test length and classification accuracy for true ability close to each classification bound when items are selected with Fisher information at $\hat{\theta}_i$ and there is no item exposure control. The column in the middle of each block is close to a classification bound. Based on Table 4.1, all of the stochastic curtailment conditions generally result in shorter tests near the classification bounds. SCSPRT must reduce the number of items relative to SPRT when the width of the indifference regions are identical, but aside from a few exceptions, as long as $\gamma < .68$, stochastic curtailment results in shorter CCTs even when the SPRT indifference region is 1.5 times as large. SPRT appears to result in shorter tests for θ_i larger than the maximum classification bound, but the relative efficiency worsens for θ_i's at higher values than included in the table (as shown in Figure 4.1). Stochastic curtailment also compares favorably to the generalized likelihood ratio test as long as the indifference regions are the same. The GLR is similar to the SPRT for θ_i's close to a classification bound, so it is not surprising that stochastic curtailment improves over the corresponding GLR stopping rule for those true ability levels. Widening the indifference region when using the GLR results in shorter tests, but as will be shown later, the gain in efficiency leads to less accurate classifications when aggregating across θ.

The counterpart to Table 4.1 is Table 4.2, which shows the percentage classified correctly at specific ability levels. Values in Table 4.2 reinforce contentions from Figure 4.3 in that most of the procedures result in a classification accuracy around 50% near each cut-point, 60%–70% at the edges of a $2\delta = .20$ indifference region, and much higher outside of the small indifference region. Inside of the indifference region, it is difficult to systematically determine which procedure performs better, and outside of the indifference region, the differences between the procedures might be more due to random noise. A final note is that stochastic curtailment with a $\gamma = .95$ correction results in the

Table 4.1: Test length averaged over 400 classification CATs conditional on particular 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 midpoint between each of the vertical bars is close to a classification bound, and only a few θ_i on either side of each classification bound are presented for clarity.

Condition	$\theta_i = -1.6$	-1.5	-1.4	-1.3	-1.2	-0.7	-0.6	-0.5	-0.4	-0.3
SPRT: $\delta = .10$	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .15$	20.5	20.6	20.8	21.0	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .20$	19.7	20.3	20.4	20.7	20.7	20.7	20.8	20.8	20.9	20.7
SPRT: $\delta = .25$	17.9	18.8	20.2	19.5	19.5	19.3	20.0	19.8	19.5	19.9
SPRT: $\delta = .30$	17.2	17.0	17.9	17.6	18.7	18.3	17.8	18.5	18.8	19.1
SCSPRT: $\delta = .20; \gamma = .95$	16.6	17.2	18.2	17.2	19.2	18.7	18.2	19.6	18.7	18.0
SCSPRT: $\delta = .20; \gamma = .86$	16.5	17.4	17.8	17.9	17.6	18.1	18.5	18.6	19.1	18.6
SCSPRT: $\delta = .20; \gamma = .68$	16.9	16.9	16.8	18.3	17.1	17.7	18.2	18.2	18.1	18.2
SCSPRT: $\delta = .20; \gamma = .38$	14.9	16.8	16.6	16.9	16.5	17.5	17.4	17.0	17.6	17.1
SCSPRT: $\delta = .20; \gamma = .00$	14.2	15.7	16.0	15.9	15.6	17.1	16.8	16.8	16.9	17.0
GLR: $\delta = .10$	18.8	19.3	19.7	20.4	20.5	20.9	21.0	21.0	21.0	21.0
GLR: $\delta = .15$	17.1	18.4	19.2	20.5	20.2	20.5	20.9	20.9	20.9	20.7
GLR: $\delta = .20$	17.4	18.1	19.1	18.8	19.5	19.9	20.3	20.4	20.4	20.6
GLR: $\delta = .25$	17.2	17.2	17.4	17.6	17.3	18.7	20.2	19.3	19.8	19.7
GLR: $\delta = .30$	16.0	16.9	17.3	17.9	17.7	17.8	18.1	18.6	18.5	18.4

Condition	$\theta_i = 0.1$	0.2	0.3	0.4	0.5	1.0	1.1	1.2	1.3	1.4
SPRT: $\delta = .10$	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .15$	21.0	21.0	21.0	20.8	20.7	20.7	20.8	20.9	20.5	20.3
SPRT: $\delta = .20$	20.7	20.6	20.0	19.6	19.2	19.3	19.5	19.0	19.2	18.7
SPRT: $\delta = .25$	19.7	19.2	19.0	18.4	18.2	17.0	17.4	17.3	17.6	16.8
SPRT: $\delta = .30$	18.0	18.0	18.1	17.4	15.5	14.5	16.6	15.9	16.0	15.3
SCSPRT: $\delta = .20; \gamma = .95$	18.1	18.7	18.3	16.2	16.7	17.1	17.4	17.5	17.6	16.7
SCSPRT: $\delta = .20; \gamma = .86$	18.7	18.6	18.5	17.7	17.6	16.8	17.1	16.8	16.8	16.0
SCSPRT: $\delta = .20; \gamma = .68$	18.2	17.9	18.2	16.9	17.3	16.1	16.5	16.8	15.7	15.3
SCSPRT: $\delta = .20; \gamma = .38$	17.6	18.6	17.6	15.9	16.2	15.6	15.2	16.1	15.6	15.4
SCSPRT: $\delta = .20; \gamma = .00$	16.9	17.0	16.8	16.5	15.1	15.3	15.9	15.5	14.4	14.2
GLR: $\delta = .10$	20.6	20.8	20.8	20.5	19.9	19.4	19.4	19.1	17.6	17.6
GLR: $\delta = .15$	20.7	20.8	20.7	19.8	18.0	19.1	18.7	17.9	17.9	17.2
GLR: $\delta = .20$	20.4	19.6	19.7	19.2	18.4	17.3	18.1	18.4	17.3	15.9
GLR: $\delta = .25$	19.1	19.1	18.8	17.3	17.4	15.8	16.7	16.3	16.3	14.6
GLR: $\delta = .30$	18.0	17.8	17.4	16.9	16.1	14.9	16.2	15.6	15.9	14.9

Table 4.2: Classification accuracy averaged over 400 classification CATs conditional on particular 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 midpoint between each of the vertical bars is close to a classification bound, and only a few θ_i on either side of each classification bound are presented for clarity.

Condition	$\theta_i = -1.6$	-1.5	-1.4	-1.3	-1.2	-0.7	-0.6	-0.5	-0.4	-0.3
SPRT: $\delta = .10$	.77	.64	.50	.64	.74	.83	.68	.60	.62	.68
SPRT: $\delta = .15$	.76	.65	.53	.64	.82	.80	.60	.56	.62	.80
SPRT: $\delta = .20$	.72	.62	.44	.73	.76	.80	.73	.49	.64	.73
SPRT: $\delta = .25$	.83	.65	.48	.61	.70	.79	.59	.56	.59	.68
SPRT: $\delta = .30$	.70	.60	.50	.65	.74	.78	.65	.46	.61	.73
SCSPRT: $\delta = .20; \gamma = .95$	.87	.50	.39	.67	.88	.70	.84	.39	.64	.78
SCSPRT: $\delta = .20; \gamma = .86$	.83	.63	.54	.56	.75	.80	.70	.53	.52	.78
SCSPRT: $\delta = .20; \gamma = .68$	.78	.64	.47	.64	.74	.80	.62	.47	.60	.70
SCSPRT: $\delta = .20; \gamma = .38$	.90	.54	.44	.70	.70	.85	.71	.46	.77	.89
SCSPRT: $\delta = .20; \gamma = .00$	.82	.57	.64	.68	.64	.70	.70	.56	.58	.79
GLR: $\delta = .10$	.70	.51	.52	.70	.71	.75	.68	.52	.63	.79
GLR: $\delta = .15$	.75	.65	.52	.75	.73	.85	.62	.46	.52	.74
GLR: $\delta = .20$	.79	.67	.54	.63	.73	.80	.69	.57	.60	.72
GLR: $\delta = .25$	.78	.75	.69	.63	.70	.91	.60	.45	.75	.69
GLR: $\delta = .30$	.75	.67	.53	.63	.70	.80	.68	.59	.62	.70

Condition	$\theta_i = 0.1$	0.2	0.3	0.4	0.5	1.0	1.1	1.2	1.3	1.4
SPRT: $\delta = .10$	.80	.62	.57	.68	.85	.79	.65	.58	.77	.83
SPRT: $\delta = .15$	.72	.60	.52	.74	.83	.75	.61	.54	.79	.87
SPRT: $\delta = .20$	.76	.62	.51	.72	.89	.80	.66	.54	.70	.85
SPRT: $\delta = .25$	.82	.58	.58	.70	.76	.70	.58	.55	.69	.85
SPRT: $\delta = .30$	.79	.57	.53	.71	.83	.72	.64	.55	.64	.84
SCSPRT: $\delta = .20; \gamma = .95$	.84	.83	.39	.78	.87	.79	.66	.58	.68	.78
SCSPRT: $\delta = .20; \gamma = .86$	.76	.60	.58	.74	.85	.83	.61	.56	.68	.82
SCSPRT: $\delta = .20; \gamma = .68$	.83	.64	.51	.75	.81	.73	.56	.50	.72	.83
SCSPRT: $\delta = .20; \gamma = .38$	.81	.58	.41	.77	.79	.76	.46	.49	.67	.77
SCSPRT: $\delta = .20; \gamma = .00$	.70	.60	.53	.77	.75	.75	.69	.56	.70	.83
GLR: $\delta = .10$	.71	.60	.41	.75	.86	.76	.53	.64	.71	.88
GLR: $\delta = .15$	.73	.60	.38	.81	.80	.81	.60	.54	.72	.91
GLR: $\delta = .20$	.77	.62	.61	.65	.78	.78	.58	.58	.72	.91
GLR: $\delta = .25$	.75	.66	.56	.79	.81	.72	.75	.57	.66	.97
GLR: $\delta = .30$	.74	.60	.49	.67	.87	.81	.63	.51	.72	.86

worst classification accuracy for θ_i's abutting each cut-point, which is most likely due to an over-correction of the maximum likelihood estimate[1]. Even though there appears to be very little difference in classification accuracy for individual ability irrespective of stopping rule, the small differences might add up when averaging across a realistic distribution of simulees.

4.2 Results 2: Aggregated across a Distribution

Most measurement specialists are interested in the accuracy and expected test length of a procedure conditional on particular values of θ_i. For example, given a specific true score, how quickly will an examinee be classified, and will the classification be correct? On the other hand, test developers and practitioners need to know the average test length and accuracy across all examinees. If specific values of θ_i seldom occur, then test length and accuracy conditional on those abilities are useless in policy considerations. To determine the overall classification accuracy and test length, 5000 $\theta \sim N(0, 1)$ were simulated, and the entire set of simulees completed each combination of conditions. The standard normal distribution was chosen as an approximation to a prior, empirical distribution for the exam that was used.

Figure 4.5 displays the average test length and classification accuracy for the 5000 simulees within each of the termination conditions. To construct Figure 4.5, results were aggregated across each of the item selection and ability estimation conditions. The left side of Figure 4.5 shows results for the SPRT and GLR conditions, whereas the right side of Figure 4.5 shows results for the SCSPRT conditions. The results were divided into two disparate plots to make the point cloud easier to parse.

Notice how termination condition relates to classification accuracy only through test

[1] Figures of conditional accuracy and average test length for the other conditions are presented in Appendix C, and the corresponding tables are presented in Appendix A.

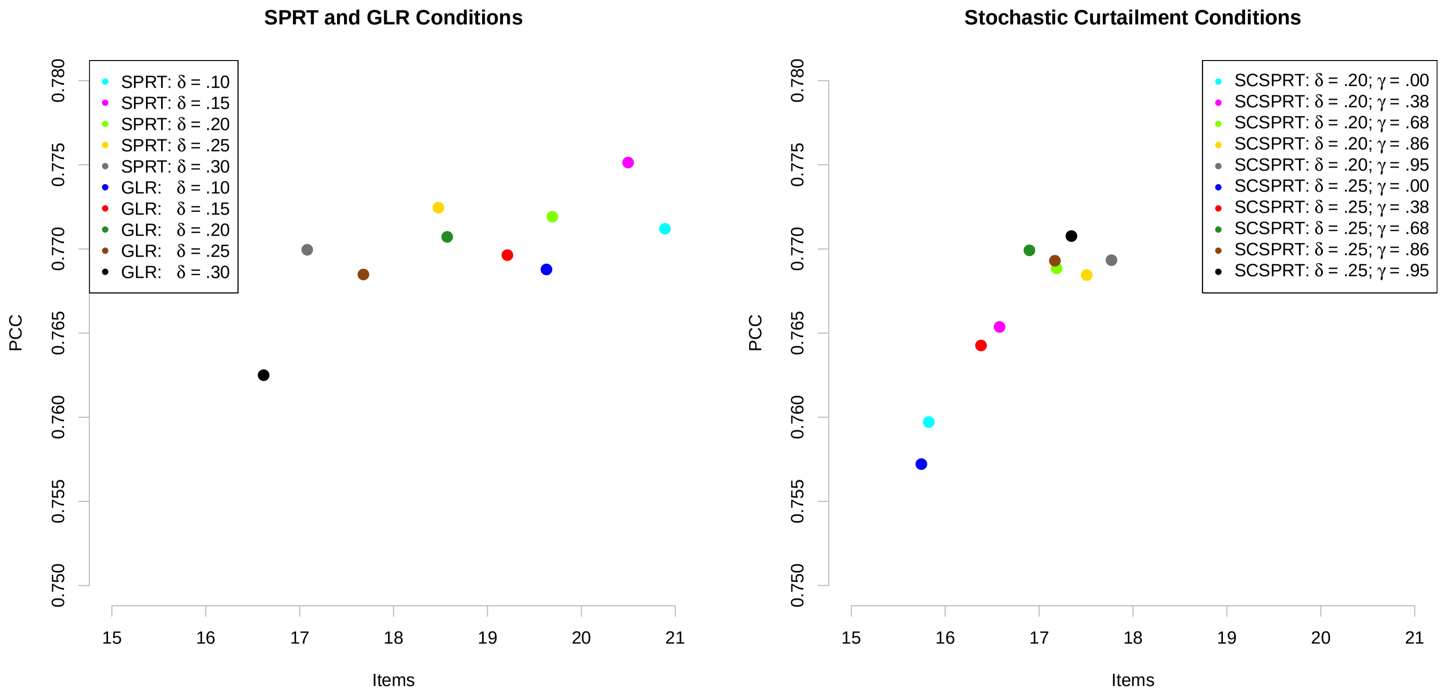


Figure 4.5: Side-by-side scatterplots of the average percentage classified correctly by number of items administered based on each termination criterion. The left plot contains all of the SPRT and GLR conditions, and the right plot contains all of the SCSPRT conditions. Points are color coded according to termination condition.

length. When a CCT is classified in 17 or more items, the mean PCC is near .770 with little relationship between test length and classification accuracy, whereas when a CTT is classified in fewer than 17 items, there appears to be a much stronger relationship between test length and classification accuracy. In fact, for those conditions with a test length of 17 or more items, the OLS slope of classification accuracy regressed on test length is $\frac{1}{10}$th the size of the corresponding OLS slope based on those conditions with a test length of less than 17 items ($\hat{\beta}_1 = .00085$ as compared to $\hat{\beta}_1 = .0089$). Yet, although this pattern holds in both plots, stochastic curtailment typically results in shorter exams for all conditions. If the confidence interval correction is wide enough ($\gamma \geq .68$), then the average percentage classified correctly using SCSPRT is at least as high as the other termination conditions. The values contained in Figure 4.5 are explicitly presented in Table 4.3. Notice how the first five rows of Table 4.3 have an average test length of fewer than 17 items and an average percent classified correctly of smaller than .765. Alternatively, aside from one row of the table, the remaining termination conditions are within .4% of each other. Notwithstanding the apparently small differences between termination criteria, Figure 4.5 only examines stopping rules by averaging across all associated conditions. One should gain a better understanding of whether certain stopping rules affect overall classification accuracy by inspecting individual conditions.

Figure 4.6 displays the average test length and classification accuracy within all combinations of conditions but color coded according to termination procedure. In a pilot study, the plot corresponding to Figure 4.6 was much clearer, and all of the conditions were well separated (Nydick, Nozawa, & Zhu, 2012). Part of the previous plot was misleading, as it was earlier claimed that stochastic curtailment resulted in reducing the number of items without much of a decrement in classification accuracy. By extending the simulation to include more termination conditions, the relationship

Table 4.3: The average test length and accuracy for all termination conditions averaged across item selection and ability estimation methods.

Condition		Mean Test Length	Mean PCC
SCSPRT:	$\delta = .25$; $\gamma = .00$	15.7	.757
SCSPRT:	$\delta = .20$; $\gamma = .00$	15.8	.760
SCSPRT:	$\delta = .25$; $\gamma = .38$	16.4	.764
SCSPRT:	$\delta = .20$; $\gamma = .38$	16.6	.765
GLR:	$\delta = .30$	16.6	.763
SCSPRT:	$\delta = .25$; $\gamma = .68$	16.9	.770
SPRT:	$\delta = .30$	17.1	.770
SCSPRT:	$\delta = .25$; $\gamma = .86$	17.2	.769
SCSPRT:	$\delta = .20$; $\gamma = .68$	17.2	.769
SCSPRT:	$\delta = .25$; $\gamma = .95$	17.3	.771
SCSPRT:	$\delta = .20$; $\gamma = .68$	17.5	.768
GLR:	$\delta = .25$	17.7	.768
SCSPRT:	$\delta = .20$; $\gamma = .95$	17.8	.769
SPRT:	$\delta = .25$	18.5	.772
GLR:	$\delta = .20$	18.6	.771
GLR:	$\delta = .15$	19.2	.770
GLR:	$\delta = .10$	19.6	.769
SPRT:	$\delta = .20$	19.7	.772
SPRT:	$\delta = .15$	20.5	.775
SPRT:	$\delta = .10$	20.9	.771

obvious in the previous study is not as apparent. The GLR conditions appear to classify simulees quicker and slightly worse than the corresponding SPRT conditions, most of the SCSPRT conditions appear to perform similarly, and as long as the test length exceeds 17 items, all of the point clouds look alike.

Much of the muddled relationship among the termination conditions is due to item exposure control. Three exposure control conditions were chosen: $r_{\max} = 1$ (every item will be administered if it is selected), $r_{\max} = .2$ (the administration of any item should be at most in $\frac{2}{10}$th of the total number of tests), and $r_{\max} = .1$ (the administration of any item should be at most in $\frac{1}{10}$th of the total number of tests). To visualize the effect of exposure control on CCT performance, Figure 4.7 displays the average test length and classification accuracy within all combinations of conditions but color coded according to exposure control condition, and Figure 4.8 presents the average test length and classification accuracy conditional on each exposure control condition and color coded according to termination procedure. The top row of Figure 4.8 includes all of the conditions such that $r_{\max} = 1$; the middle row of Figure 4.8 includes all of the conditions such that $r_{\max} = .2$; and the bottom row of Figure 4.8 includes all of the conditions such that $r_{\max} = .1$.

Notice how the green scatter of Figure 4.7 (representing those conditions with the most stringent exposure control) is below and to the right of the blue and pink scatter. Because a stringent exposure control tends to reduce classification accuracy and increase test length regardless of termination condition, the color-coded point clouds of Figure 4.6 are diagonal and appear to greatly overlap. After conditioning on exposure control condition, much of the overlap disappears. When there is no overlap (the top row of Figure 4.8), the optimal condition (shortest average test lengths with relatively highest classification accuracy) appears to be SPRT with $\delta = .30$. As long as there is no exposure control, increasing the half-width of the indifference region results in much

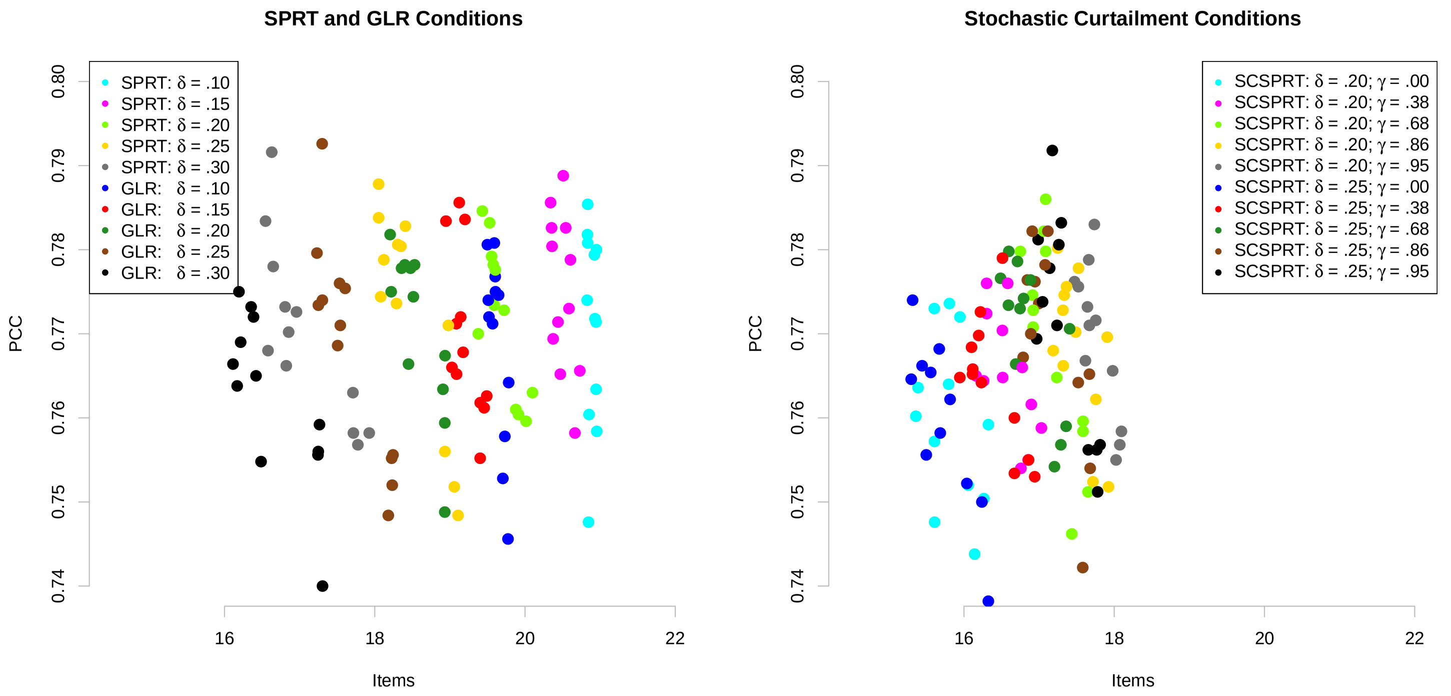


Figure 4.6: Side-by-side scatterplots of the percentage classified correctly by number of items administered based on each termination criterion. The left plot contains all of the SPRT and GLR conditions, and the right plot contains all of the SCSPRT conditions. Points are color coded according to termination condition, and differing points within a particular color represent a combination of item selection condition, ability estimation, and exposure control conditions.

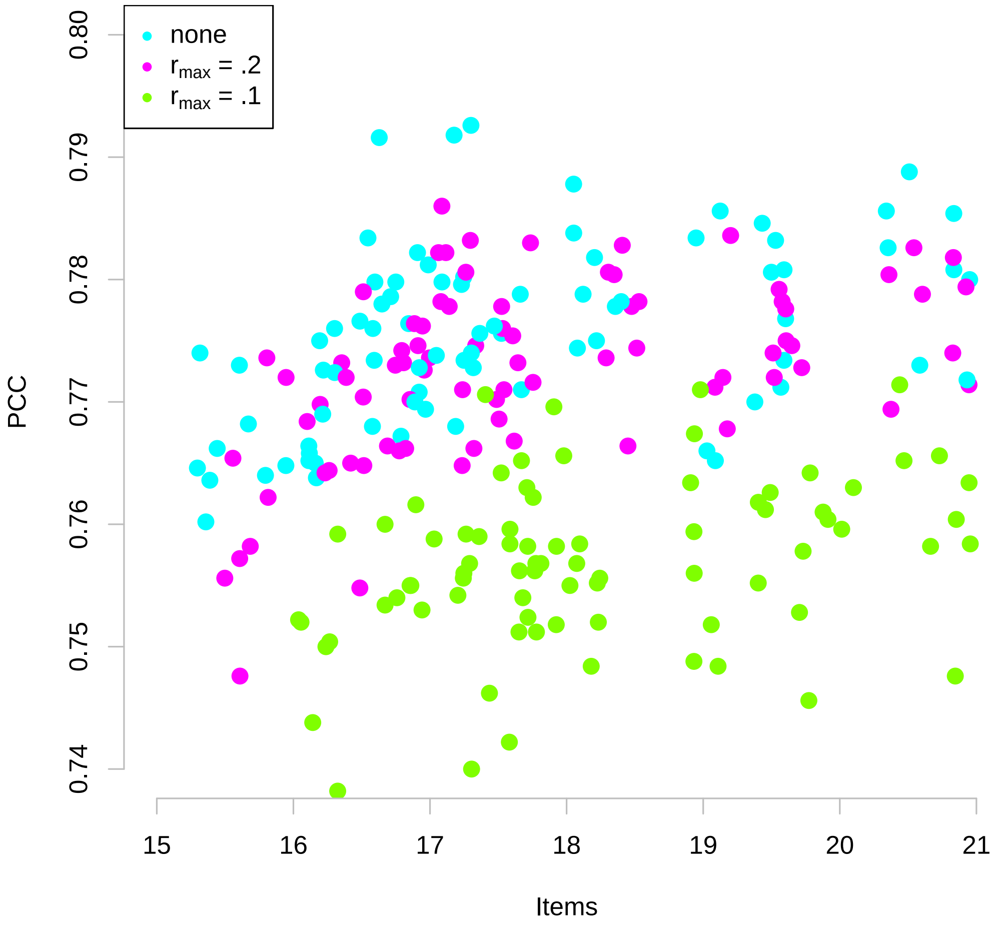
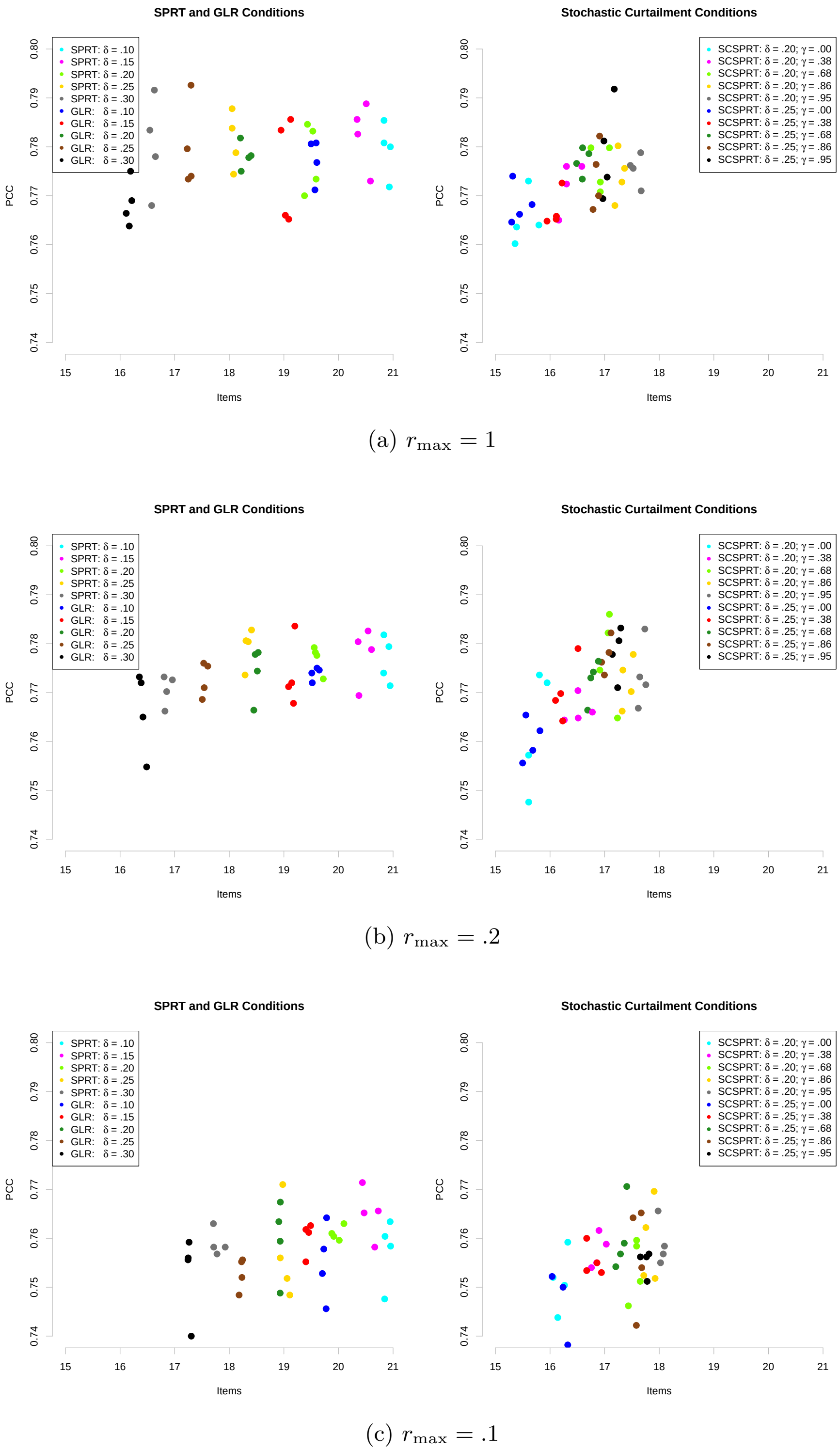


Figure 4.7: A scatterplot of the percentage classified correctly by number of items administered based on each exposure control condition. Points are color coded according to exposure control condition ($r_{\max} = 1$, .2, or .1), and differing points within a particular color represent a combination of item selection condition, ability estimation, and termination conditions.



(a) $r_{\max} = 1$

(b) $r_{\max} = .2$

(c) $r_{\max} = .1$

Figure 4.8: Side-by-side scatterplots of the average percentage classified correctly by number of items administered based on each termination criterion within a given exposure control condition. Points are color coded according to termination condition.

shorter CCTs but does not appear to affect classification accuracy. After implementing an item exposure correction, the relative performance of the SCSPRT conditions improve. The SCSPRT point clouds in the middle set of plots are slightly higher than the corresponding SPRT/GLR point clouds[2].

To reinforce conclusions drawn from Figures 4.5–4.6, two ANOVA table were constructed: (1) Table 4.4 summarizes the effect of different factors on mean test length; and (2) Table 4.5 summarizes the effect of the same factors on mean classification accuracy. Only main effects and those two-way interactions that include the stopping rule factor were examined. Using an ANOVA is arguably inappropriate considering the dependent variable and research design, but it is only being used as a descriptive measure of variance accounted for by each factor (e.g., Guyer and Weiss, 2009). Tables 4.4 and 4.5 both present $\eta^2 = \frac{SSF}{SST}$, where SSF is the sums of squares of a particular factor and SST is the total sums of squares. Similar to multiple R^2 in linear regression, η^2 is a positively biased estimate of the variance accounted for by each factor in the population. Additionally, ω^2 was also calculated, and it is a less biased estimate of the proportion of variance (e.g., Olejnik & Algina, 2000), but the substantive conclusions would have been identical, and I chose to present η^2 due to its straightforward interpretation. Based on Tables 4.4 and 4.5, the stopping rule, item exposure, and the interaction between the stopping rule and item exposure accounts for most of the variance of both percentage classified correctly and test length. Stopping rule accounts for most of the variance in test length ($\eta^2 = .958$), which is not surprising considering that conditions were chosen that are known to affect test length, and exposure control accounts for most of the remaining variance in test length ($\eta^2 = .032$). Item selection, ability estimation, and all of the interactions account for so little variability in test length that those conditions

[2] Figures displaying the relationship between average test length and classification accuracy for a variety of conditions are presented in Appendix D, and the corresponding main effects tables are shown in Appendix B.

are not worth discussing. Examining Table 4.5, exposure control accounts for most of the observable variance in classification accuracy ($\eta^2 = .568$), although there is a small effect of termination criterion ($\eta^2 = .148$) and the interaction between exposure control and termination ($\eta^2 = .041$). It is not surprising that termination condition accounts for some of the variance in classification accuracy because most of the rules that tend to terminate before 17 items result in a large drop in average percent classified correctly (PCC). After removing those conditions in which the average test length is fewer than 17 items, the variance accounted for by stopping rule drops to $\eta^2 = .028$ and the remaining proportion of variance redistributes to the item exposure factor (see Table B.5 in Appendix B). The main effects of item selection, ability estimation, and the remaining two-way interactions account for so little variance in classification accuracy that I chose not to discuss them. Finally, the interaction between exposure control and termination condition had a non-trivial proportion of variance ($\eta^2 = .041$), which was alluded to earlier, although future studies should address whether the non-trivial interaction is due to the conditions examined in this study.

A problem encountered when running simulations was an inability for the SPRT to classify certain simulees. As was alluded to in Section 4.1, the SPRT took longer to classify examinees above the maximum cut-point than below the minimum cut-point. Because the relative weaknesses of the SPRT are due to properties of the likelihood ratio test statistic, I now explain why the SPRT is inefficient as an illumination of how test developers should design item banks and choose classification criteria for adaptive tests.

Table 4.4: The sums of squares and $\eta^2 = \frac{SSF}{SST}$, where SSF is the sums of squares of a particular factor, for an ANOVA predicting mean test length. The ANOVA was run with all main effects and those interactions that relate to the termination factor.

Variance Type	Sums of Squares	η^2
Termination	506.50	.958
Exposure	16.80	.032
Term by Expos	3.73	.007
Term by Estim	0.67	.001
Term by Select	0.58	.001
Ability Estimation	0.27	.001
Item Selection	0.01	.000
Residuals	0.27	
Total	528.84	

Table 4.5: The sums of squares and $\eta^2 = \frac{SSF}{SST}$, where SSF is the sums of squares of a particular factor, for an ANOVA predicting mean classification accuracy. The ANOVA was run with all main effects and those interactions that relate to the termination factor.

Variance Type	Sums of Squares	η^2
Exposure	0.01666	.568
Termination	0.00434	.148
Term by Expos	0.00121	.041
Term by Estim	0.00061	.021
Item Selection	0.00057	.019
Term by Select	0.00046	.016
Ability Estimation	0.00043	.015
Residuals	0.00505	
Total	0.02933	

4.3 SPRT and the Three-Parameter Logistic Model

Practitioners should exercise caution when using the three-parameter logistic model in classification CAT, especially when classifying examinees using the SPRT. Several researchers have alluded to limitations of the 3PL when examining other IRT models. For example Finkelman, Hooker, and Wang (2009) noticed that one of two sufficient conditions for avoiding paradoxical results[3] is that "all second derivatives of the log of the item response surface should be strictly negative" (p. 420), a trait that the 3PL does not share. However, the series of articles by Hooker and colleagues (Finkelman, Hooker, & Wang, 2009; Hooker, 2010; Hooker & Finkelman, 2010; Hooker, Finkelman, & Schwartzman, 2009) was devoted to diagnosing and eschewing paradoxical results in multidimensional models, and they did not further discuss their results in a three-parameter IRT model context.

Unfortunately, many researchers do not appear knowledgeable of the relationship between the 3PL and test length. Thompson (2010), who used the 3PL in his simulations, thought that "it is far easier to make a classification if the cut-score is in the extremes" and that "only a few items might be needed to classify an examinee above a cut-score of -2.0 or below $+2.0$" (p. 9), a characteristic that is not true in models with nonzero, lower asymptotes. Only Spray and Reckase (1994) explicitly acknowledged that "the large difference in number of items for the high ability examinees is a result of the nonzero lower asymptote for the three parameter logistic model" (p. 7). Because the 3PL is popular in educational testing, and in view of the fact that classification researchers do not appear to be aware of its limitations in CCT, I briefly illustrate the consequence of a nonzero lower asymptote on the likelihood ratio test statistic.

Without loss of generality, consider a classification task involving one cut-point and

[3] Paradoxical results in multidimensional IRT is a phenomenon whereby answering a question correctly can lower an ability estimate on at least one dimension.

a symmetric indifference region of size 2δ. Conceiving the likelihood ratio test statistic as a function of the classification bound, θ_0, Equation (2.3) for examinee i after item J can be written

$$\begin{aligned}\log\left[LR(\theta_0+\delta,\theta_0-\delta|\mathbf{u}_i)\right] &= \log\left[\frac{L(\theta_0+\delta|\mathbf{u}_i)}{L(\theta_0-\delta|\mathbf{u}_i)}\right]\\ &= \sum_{j=1}^{J} u_{ij}\log\left[\frac{p_j(\theta_0+\delta)}{p_j(\theta_0-\delta)}\right]+\sum_{j=1}^{J}(1-u_{ij})\log\left[\frac{1-p_j(\theta_0+\delta)}{1-p_j(\theta_0-\delta)}\right].\end{aligned} \tag{4.1}$$

Spray and Reckase (1994) noticed that when $c_j > 0$ then $\lim_{\theta_0\to-\infty}\frac{p_j(\theta_0+\delta)}{p_j(\theta_0-\delta)} = 1$ and $\lim_{\theta_0\to-\infty}\frac{1-p_j(\theta_0+\delta)}{1-p_j(\theta_0-\delta)} = 1$. Therefore, when all of the pseudo-guessing parameters are greater than 0 and the classification bound is extremely negative, Equation (4.1) will be close to 0 regardless of an examinee's true ability. Taking the first derivative of Equation (4.1) results in

$$\frac{d\log\left[LR(\theta_0+\delta,\theta_0-\delta|\mathbf{u}_i)\right]}{d\theta_0} = \sum_{j=1}^{J} a_j u_{ij}[p_j^{c_j}(\theta_0+\delta)-p_j^{c_j}(\theta_0-\delta)]-\sum_{j=1}^{J} a_j[p_j^{1}(\theta_0+\delta)-p_j^{1}(\theta_0-\delta)], \tag{4.2}$$

where $p_j^{c_j}(\theta_0) = \frac{\exp[a_j(\theta_0-b_j)]}{c_j+\exp[a_j(\theta_0-b_j)]}$ and $p_j^{1}(\theta_0) = \frac{\exp[a_j(\theta_0-b_j)]}{1+\exp[a_j(\theta_0-b_j)]}$, as defined in Equation (2.10). In a model without c parameters, $p_j^{c_j}(\theta_0+\delta) - p_j^{c_j}(\theta_0-\delta) = 1 - 1 = 0$ for all items, so that

$$\frac{d\log\left[LR(\theta_0+\delta,\theta_0-\delta|\mathbf{u}_i)\right]}{d\theta_0} = -\sum_{j=1}^{J} a_j[p_j^{1}(\theta_0+\delta)-p_j^{1}(\theta_0-\delta)], \tag{4.3}$$

which does not depend on an examinee's item responses. Importantly, the sign of Equation (4.3) is always negative (unless $\delta = 0$), so that the log-likelihood ratio test statistic is monotonically decreasing as the classification bound increases. The consequence of a monotonic log-likelihood ratio statistic can be explained with a simple example. Pretend that $\theta_i = 2$ and there are two classification bounds of $\theta_1 = 0$ and $\theta_2 = 1$. When

$c_j = 0$ for all items on an exam, then the log-likelihood ratio test statistic will be larger comparing $\theta_1 + \delta$ to $\theta_1 - \delta$ than comparing $\theta_2 + \delta$ to $\theta_2 - \delta$, providing more evidence that examinee i is above $\theta_1 = 0$ than $\theta_2 = 1$. When there are no pseudo-guessing parameters, Thompson's (2010) assertion that "it is far easier to make a classification if the cut-score is in the extremes" (p. 9) is accurate.

However, when any $c_j > 0$, then the log-likelihood ratio is not necessarily monotonic. As a demonstration, take expectations of Equation (4.2) to remove the individual response pattern, and set the derivative equal to zero, which results in

$$\begin{aligned}\sum_{j=1}^{J} a_j [p_j^1(\theta_0 + \delta) - p_j^1(\theta_0 - \delta)] &= \sum_{j=1}^{J} a_j p_j(\theta_i)[p_j^{c_j}(\theta_0 + \delta) - p_j^{c_j}(\theta_0 - \delta)], \\ &= \sum_{j=1}^{J} a_j \left[\left(\frac{p_j(\theta_i)}{p_j(\theta_0 + \delta)} \right) p_j^1(\theta_0 + \delta) - \left(\frac{p_j(\theta_i)}{p_j(\theta_0 - \delta)} \right) p_j^1(\theta_0 - \delta) \right], \\ \sum_{j=1}^{J} a_j p_j^1(\theta_0 + \delta) \left[1 - \frac{p_j(\theta_i)}{p_j(\theta_0 + \delta)} \right] &= \sum_{j=1}^{J} a_j p_j^1(\theta_0 - \delta) \left[1 - \frac{p_j(\theta_i)}{p_j(\theta_0 - \delta)} \right]. \qquad (4.4)\end{aligned}$$

The relationship between θ_0, δ, θ_i, the item parameters, and Equation (4.4) being satisfied is complex. The illustration can be simplified by assuming that there is only one item with $a = 1$ and $b = 0$. Given that item, each half of Equation (4.4) can be calculated for various values of c, θ_0, δ, and θ_i. Note that it is only necessary to test $\theta_i \notin (\theta_0 - \delta, \theta_0 + \delta)$ because when θ_i is within the indifference region, then both halves of Equation (4.4) are necessarily of different sign. Now let $c = .2$, $\delta = .1$, $\theta_i = 2$, and vary θ_0 from -4 to 2. Then each half of Equation (4.4) is displayed on the left side of Figure 4.9, and the full expected derivative is displayed on the right side of Figure 4.9. When $\theta_0 < -.94$, then $p^1(\theta_0 + \delta)\left[1 - \frac{p(\theta_i)}{p(\theta_0+\delta)}\right] < p^1(\theta_0 - \delta)\left[1 - \frac{p(\theta_i)}{p(\theta_0-\delta)}\right]$, but at approximately $\theta_0 = -.94$, the curves cross, and then $p^1(\theta_0 + \delta)\left[1 - \frac{p(\theta_i)}{p(\theta_0+\delta)}\right] > p^1(\theta_0 - \delta)\left[1 - \frac{p(\theta_i)}{p(\theta_0-\delta)}\right]$. For these sets of parameters, the strongest evidence for $\theta_i > \theta_0$ is when $\theta_0 \approx -.94$ and not when $\theta_0 < -2$. In fact, when $\theta_0 = -4$, the log-likelihood ratio is approximately

.013, whereas when $\theta_0 = .94$, the log-likelihood ratio is approximately .076, providing six times the evidence that θ_i is in the upper category.

It can also be determined whether changing δ, c, and θ_i interacts with conclusions drawn from the previous plot. Figure 4.10 displays three sets of plots, each identically constructed to Figure 4.9, only with differing values of δ. In the upper third of Figure 4.10, $\delta = .05$, in the middle third, $\delta = .1$, and in the lower third, $\delta = .2$. Notice that increasing δ (i.e., changing the width of the indifference region) only increases the distance between the curves. As in Figure 4.9, when $\theta_0 \approx -.94$, the log-likelihood ratio is at a maximum, and any classification bound below $-.94$ is associated with less evidence that $\theta_i = 2$ is above it.

Figure 4.11 also displays three sets of plots, each identically constructed to Figures 4.9 and 4.10, only with differing values of c. In the upper third of Figure 4.11, $c = 0$ (i.e., no pseudo-guessing parameter), in the middle third, $c = .01$, and in the lower third, $c = .1$. Unlike altering δ, increasing the pseudo-guessing parameter changes the latent trait value where the derivative of the log-likelihood ratio is expected to equal 0. When $c = 0$, the blue curve is always greater than the red curve, so that the expected derivative of the log-likelihood ratio is always negative, which supports the conclusion drawn from Equation (4.3). However, as c increases, the intersection of the blue and red curves approaches a value slightly below $b = 0$. Below the intersection point, the change in the log-likelihood ratio is contrary to intuition in that as the classification bound approaches negative infinity, the log-likelihood ratio decreases, providing *less* evidence that an examinee with high ability is above it. Because the c-parameter changes the location of the intersection point, increasing c decreases the distance between the location of an item and the area where the log-likelihood ratio behaves counterintuitively.

Finally, Figure 4.12 displays three sets of plots with θ_0 on the x-axis and different values of θ_i used to construct each row. In the upper third of Figure 4.12, $\theta_i = b = 0$, in

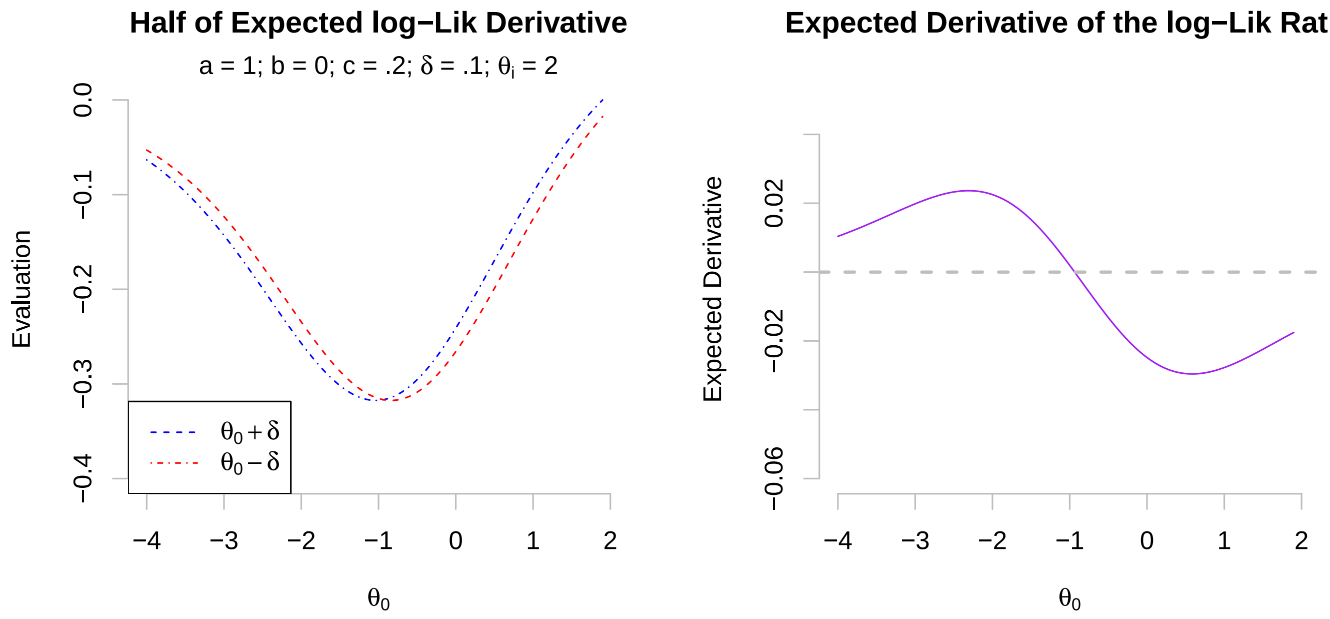


Figure 4.9: The left plot shows each half of the expected derivative of the log-likelihood ratio test statistic when $a = 1$, $b = 0$, $c = .2$, $\theta_i = 2$, $\delta = .1$, and θ_0 is varied from -4 to 2 as displayed in Equation (4.4). The right plot shows the full expected derivative as presented in Equation (4.2).

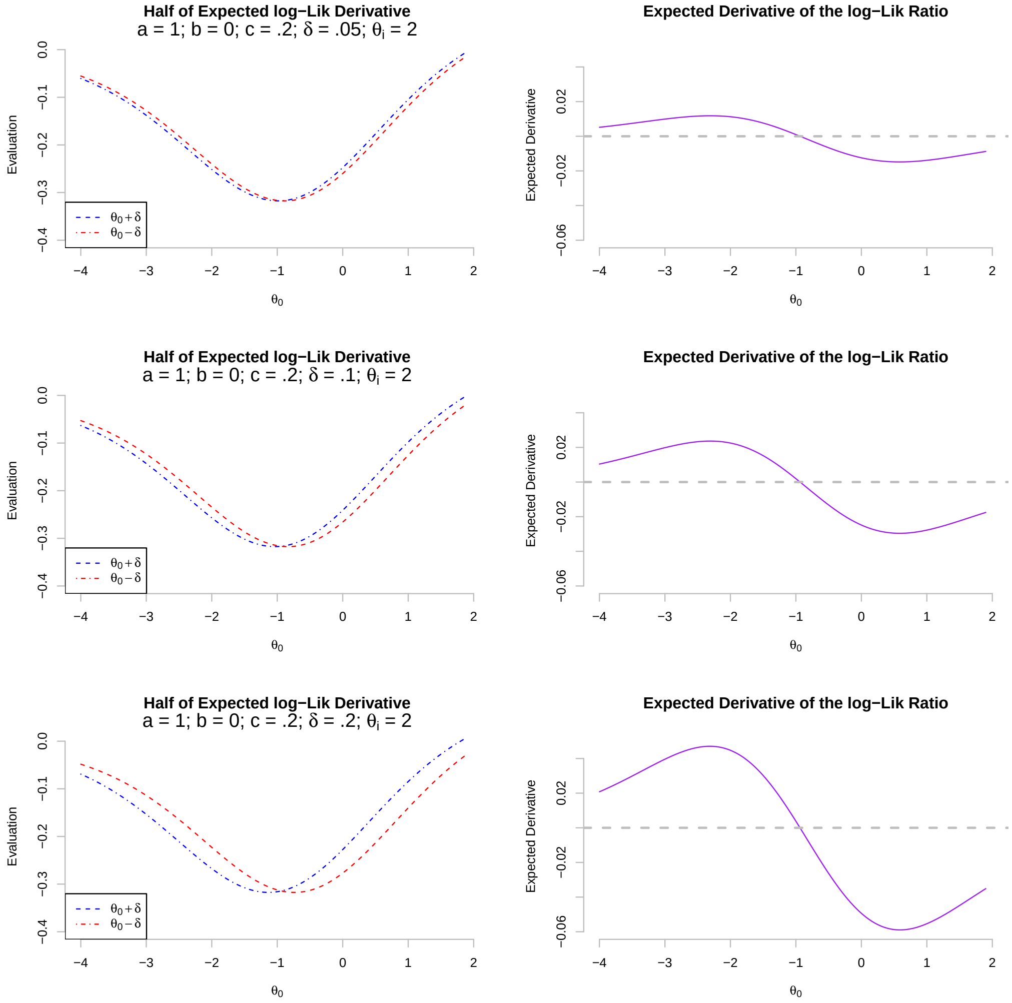


Figure 4.10: The left plots show each half of the expected derivative of the log-likelihood ratio test statistic when $a = 1$, $b = 0$, $c = .2$, $\theta_i = 2$, and θ_0 is varied from -4 to 2 as displayed in Equation (4.4). The right plots show the full expected derivative as presented in Equation (4.2). In the top plots, $\delta = .05$, in the middle plots, $\delta = .1$, and in the bottom plots, $\delta = .2$

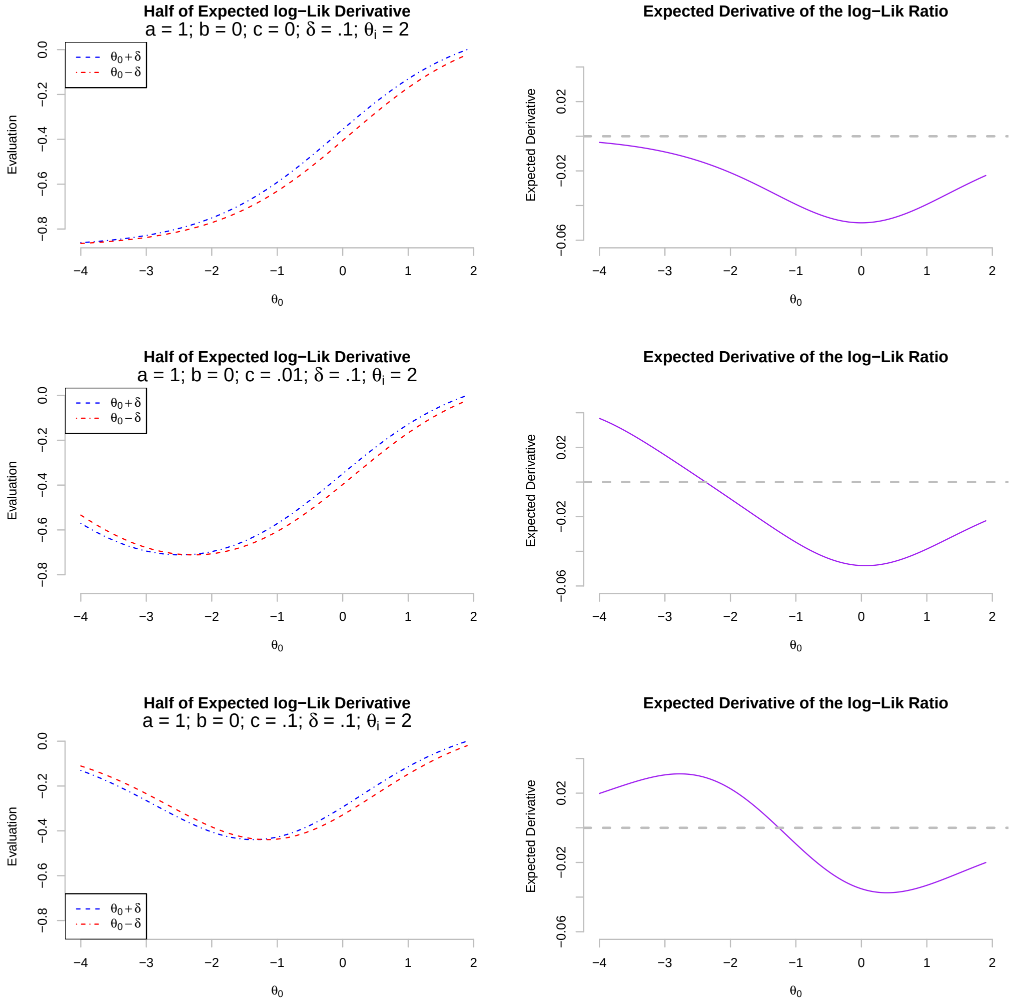


Figure 4.11: The left plots show each half of the expected derivative of the log-likelihood ratio test statistic when $a = 1$, $b = 0$, $\theta_i = 2$, $\delta = .1$, and θ_0 is varied from -4 to 2 as displayed in Equation (4.4). The right plots show the full expected derivative as presented in Equation (4.2). In the top plots, $c = 0$, in the middle plots, $c = .01$, and in the bottom plots, $c = .1$

the middle third, $\theta_i = 1$, and in the lower third, $\theta_i = 2$. It appears as though decreasing θ_i slightly decreases the location of the intersection point and also decreases the distance between the curves below the intersection point. Because for $c > 0$, the expected log-likelihood ratio approaches 0 as the classification bound approaches negative infinity, the decreased distance between the curves implies a smaller peak of the log-likelihood ratio function.

The relationship between θ_0 and the log-likelihood ratio test statistic can also be depicted when there are multiple items. To see how non-zero lower asymptotes affect classification efficiency, a 100 item CAT was simulated for $\theta_i = -3.1$ and $\theta_i = 2.9$. Following the CAT, an alternative, two-parameter dataset was created by taking the items administered to each examinee and removing all of the c-parameters. Finally, the log-likelihood ratio was calculated based on the original, three-parameter dataset and based on the modified, two-parameter dataset for various values of θ_0. The plots are displayed in Figures 4.13 and 4.14. Notice how the log-likelihood ratio is monotonically decreasing when responses follow the 2PL. As the strength of evidence decreases, the ability to classify a simulee in the upper category also decreases. And in the 2PL, an examinee classified above $\theta_0 = 0$ will also be classified above any $\theta_0 < 0$ provided that δ, α, and β remain the same. Unfortunately, when there are non-zero lower asymptotes, the relationship between θ_0 and the log-likelihood ratio is not as straightforward. For instance, consider the right plot of Figure 4.14. When $\theta_0 \approx 1$, the log-likelihood ratio is approximately 11.9 and $\theta_i = 2.9$ would be classified in the upper category. However, if $\theta_0 \approx -1$, then the log-likelihood ratio would be approximately 2.9, which is smaller than the critical value of $C_u = \log\left(\frac{1-\beta}{\alpha}\right) = \log\left(\frac{.95}{.05}\right) \approx 2.94$, and the SPRT would not be able to classify the examinee. Part of the reason that the log-likelihood ratio is so low for $\theta_0 < -1$ is because the set of items was chosen to maximize Fisher information at $\hat{\theta}_i$ and *not* at θ_0.

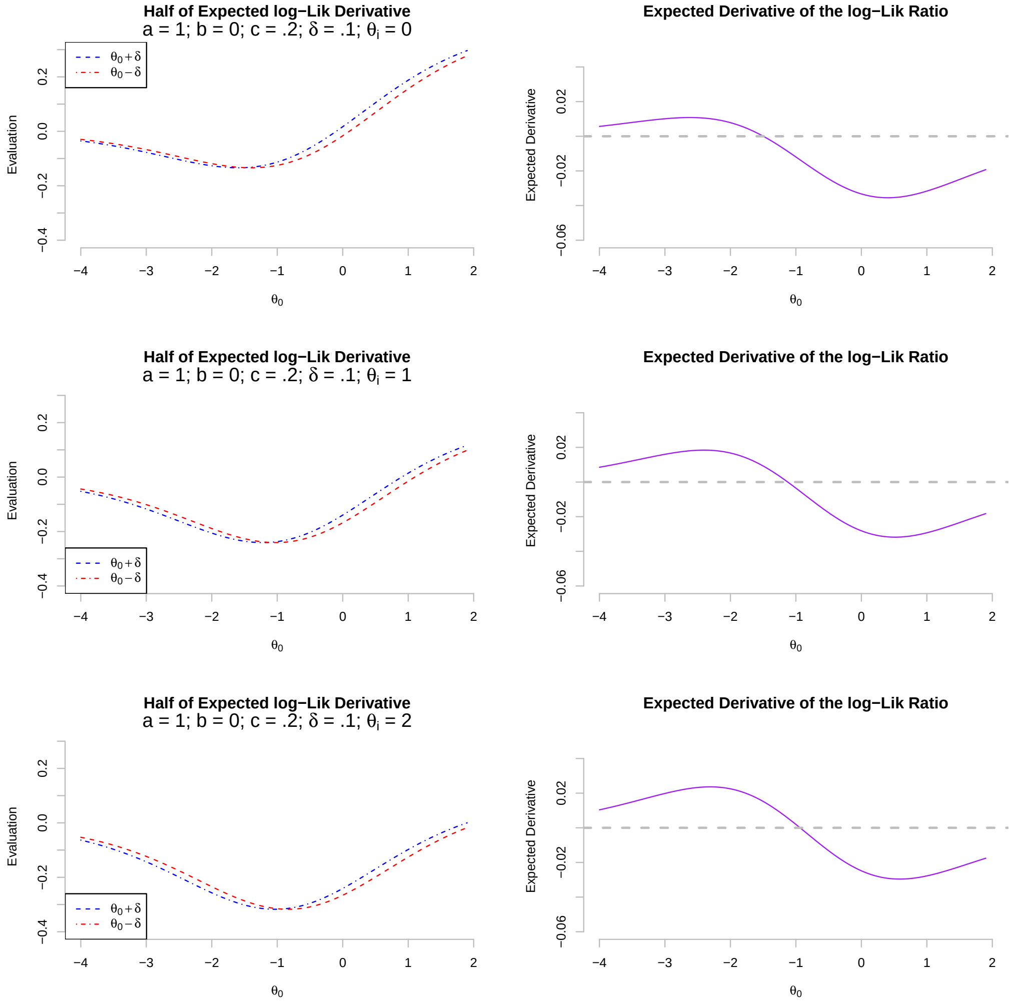


Figure 4.12: The left plots show each half of the expected derivative of the log-likelihood ratio test statistic when $a = 1$, $b = 0$, $c = .2$, $\delta = .1$, and θ_0 is varied from -4 to 2 as displayed in Equation (4.4). The right plots show the full expected derivative as presented in Equation (4.2). In the top plots, $\theta_i = 0$, in the middle plots, $\theta_i = 1$, and in the bottom plots, $\theta_i = 2$

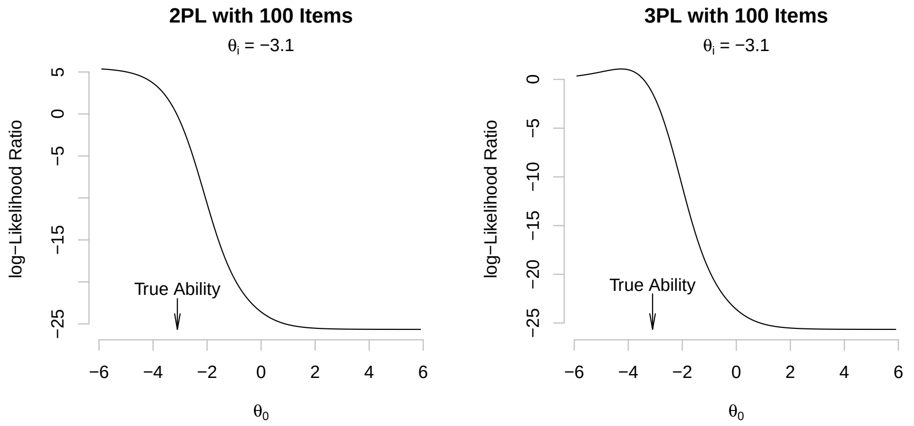


Figure 4.13: The log-likelihood ratio testing $\theta_0 + \delta$ against $\theta_0 - \delta$ with $\delta = .1$ after a 100 item CAT when $\theta_i = -3.1$. For the right plot, items were selected from the item bank calibrated under the 3PL, and for the left plot, all of the c parameters were set to 0. The x-axis indicates various values of θ_0.

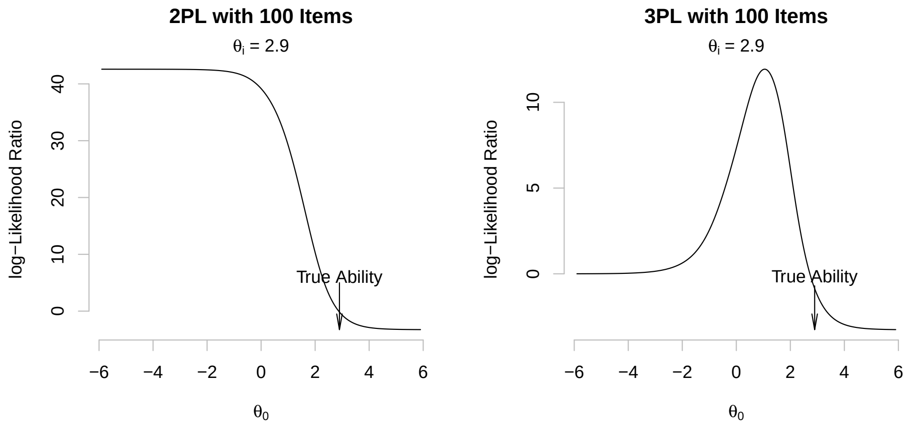


Figure 4.14: The log-likelihood ratio testing $\theta_0 + \delta$ against $\theta_0 - \delta$ with $\delta = .1$ after a 100 item CAT when $\theta_i = 2.9$. For the right plot, items were selected from the item bank calibrated under the 3PL, and for the left plot, all of the c parameters were set to 0. The x-axis indicates various values of θ_0.

Figures 4.13 and 4.14 do not reflect the change in the log-likelihood ratio for a fixed classification bound and varying true ability levels. As it turns out, when items are randomly administered to simulees, a larger true value of θ generally corresponds to a larger log-likelihood ratio in both the 2PL and the 3PL. To see the change in the log-likelihood ratio with a fixed classification bound and various ability levels, 2PL and 3PL responses were simulated to 100 randomly selected items for θ_i spaced in .01 increments from -5 to 5. Then, the log-likelihood ratio test statistic was calculated with θ_0 fixed at either 0 (as displayed in Figure 4.15) or -3 (as displayed in Figure 4.16). Irrespective of model, the log-likelihood ratio is typically higher for simulees further away from the θ_0. Yet under the 3PL, simulees above the cutting point predominantly have a lower log-likelihood ratio when $\theta_0 = -3$ as compared to $\theta_0 = 0$, whereas under the 2PL, simulees above the cutting point have a higher log-likelihood ratio. Due to properties of the 3PL, the multiple category generalization of the SPRT for IRT models (e.g., Spray, 1993) would not necessarily make consistent decisions at every bound unless examinees were administered sufficient items with difficulty levels near each of the cut-points. Therefore, practitioners who desire to use the SPRT for classification should either write items to fit the 2PL or be mindful of the relationship between the location of items and the strength of classification evidence when choosing cut-points and selecting items. For instance, when using the SPRT with the 2PL, the composition of items on a each CCT should not greatly affect test length. However, when using the SPRT with the 3PL, items should be administered with difficulty close to cut-points. Otherwise, the SPRT stopping rule will never have enough evidence to classify high ability examinees.

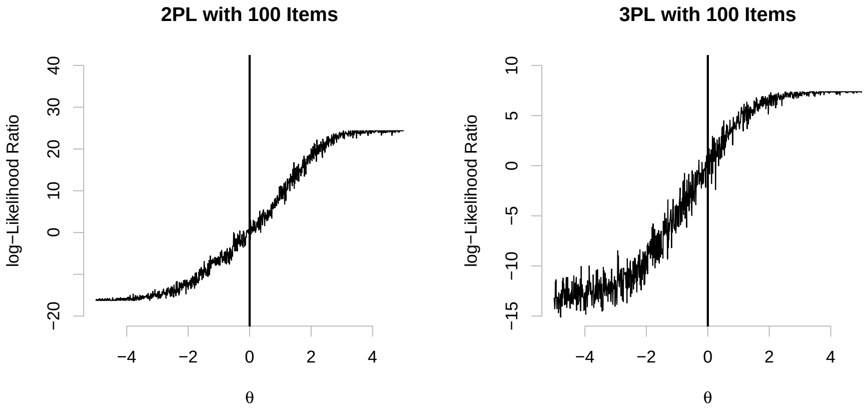


Figure 4.15: The log-likelihood ratio testing $\theta_0 + \delta$ against $\theta_0 - \delta$ with $\delta = .1$ after 100 randomly chosen and administered items. For the right plot, items were selected from the item bank calibrated under the 3PL, and for the left plot, all of the c-parameters were set to 0. The x-axis indicates various values of θ_i and the vertical line denotes a classification bound of $\theta_0 = 0$.

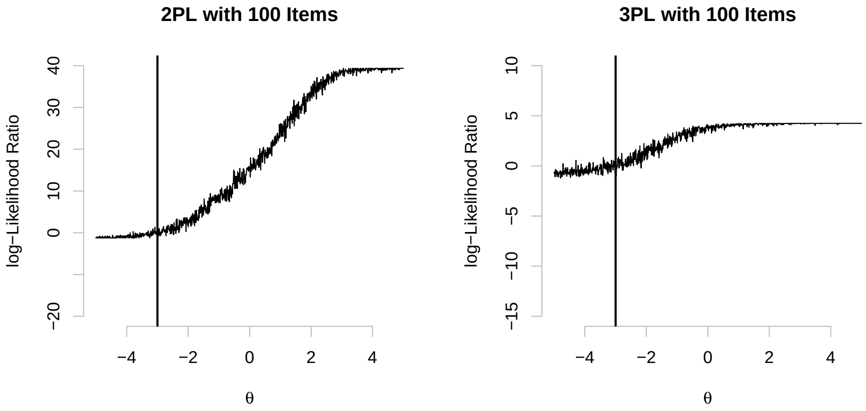


Figure 4.16: The log-likelihood ratio testing $\theta_0 + \delta$ against $\theta_0 - \delta$ with $\delta = .1$ after 100 randomly chosen and administered items. For the right plot, items were selected from the item bank calibrated under the 3PL, and for the left plot, all of the c-parameters were set to 0. The x-axis indicates various values of θ_i and the vertical line denotes a classification bound of $\theta_0 = -3$.

Chapter 5

Discussion and Conclusion

5.1 Summary and Discussion of Results

This study compared the classification accuracy and test length of various termination procedures using items culled from a real item bank. Stochastically curtailed SPRT tended to improve over the truncated SPRT in terms of test length as long as the confidence interval correction was not too large, and the accuracy trade off was acceptable as long as the confidence interval correction was not too small. On the other hand, one cannot easily compare simple hypothesis procedures with a generalized hypothesis test that takes into account the current estimate of $\hat{\theta}$. Based on the above simulations, the generalized likelihood ratio was always less accurate than an SPRT procedure that administered the same number of items. Part of the accuracy decrement could be due to using the same critical values as was used in the SPRT procedure, and a more sophisticated simulation method (e.g., Bartroff et al., 2008) might be needed to determine appropriate thresholds for classification.

Finkelman (2010) recently proposed variations on stochastic curtailment that consider 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), which was only derived for the special case of choosing one of three hypotheses about a normal distribution mean. It is clear that critical values from the typical SPRT (i.e., $C_l = \log[\beta/(1-\alpha)]$ and $C_u = \log[(1-\beta)\alpha])$ are inappropriate when there are many categories. Several researchers have proposed individual critical values based on either the distance between categories (e.g., Spray, 1993) or a step-down procedure using a rank ordering of the likelihood ratio test statistics (e.g., Bartroff & Lai, 2010; Paulson, 1963). Other researchers have extended sequential testing to multiple composite hypotheses (e.g., Pavlov, 1998), but these have yet to be applied to adaptive testing.

A promising alternative to classical stochastic curtailment is the Bayesian-like predictive power formulation (Jennison & Turnbull, 2000) applied by Finkelman (2010) to mastery tests. Predictive power weights the probability of being in a particular category at the end of the test by the posterior distribution of θ given the items already taken. Dmitrienko and Wang (2006) wrote that the predictive power formulation “has been criticized in the literature because it does not have a clear frequentist interpretation and, at the same time, is inconsistent with the principles of the Bayesian theory” (p. 2179). Fully Bayesian methods of curtailment have been implemented in the clinical trials literature (e.g., Dmitrienko & Wang, 2006), although “the choice of prior distributions can potentially have a pronounced effect on Bayesian predictive inferences” (Dmitrienko & Wang, 2006, p. 2182). In any case, more research is needed to decide whether Bayesian methods are feasible in CCT and whether modifications of Bayesian procedures improve error rates in multiple category, classification problems.

Practitioners should also bear in mind the computation time of stochastic curtailment when considering adaptive testing algorithms. Even though stochastic curtailment decreased test length without much of a reduction in classification accuracy, the computing speed was three times as long as the other two methods using R (R Development Core Team, 2011) on a 2 GHz Intel Core i7 processor. The increase in computing time is due to the SCSPRT algorithm iteratively determining the j_{remain} best items to administer before checking classification probabilities. For adaptive tests with more stringent content constraints (e.g., van der Linden, 2005), future item selection is frequently built into each state of the test, so the j_{remain} items needed to calculate components of SC-SPRT would be a by-product of the item selection algorithm.

Due to the computation time of stochastic curtailment, none of the implemented algorithms were ideal for the 3PL with multiple categories. The SCSPRT was computationally inefficient, the GLR did not have appropriate critical values, and as alluded to in Section 4.1, the SPRT took longer to classify examinees above the maximum cut-point than below the minimum cut-point. Even though no stopping rule performed as well as had been hoped, the methods described in this paper can easily be extended and examined in future research. One such study will be outlined in the next Section.

5.2 Conclusion and Proposal

Many variations of sequential tests have been developed in the statistical and biometric literature, some designed to choose between multiple composite hypotheses, and others intended to account for a maximum number of items, but few of those methods have been applied to psychometric questions. A future study should compare variations on stochastic curtailment (e.g., Dmitrienko, & Wang, 2006; Finkelman, 2010) with modifications on the original SPRT and GLR algorithms (e.g., Bartroff et al., 2008;

Bartroff & Lai, 2010; Finkelman, 2008a; Spray, 1993) that better control the desired error rates. One such method is strict stochastic curtailment (referred to as "conditional power" by Dmitrienko & Wang, 2006), but there are also "predictive power" and "predictive probability" approaches, the former adopting a mix between frequentist and Bayesian philosophies and the latter opting for a strictly Bayesian framework. Bayesian modifications of sequential testing have been adopted in classification algorithms (e.g., Kingsbury & Weiss, 1983; Lewis & Shehan, 1992; Spray & Reckase, 1996; Vos & Glas, 2010), but few curtailment methods have been adequately applied to classification CAT with more than two categories. To determine whether a proposed method works well, one also needs to vary the number of cut-points (between 1 and 3 to resemble real classification problems), the maximum test length (between 25, 50, and 100 to be similar to previous studies on stopping rule efficacy), and the IRT model (either the 2PL or the 3PL). It was already determined that a guessing parameter might affect classification criteria, so it would be valuable to build parallel item banks (i.e., those that have similar test information curves) of varying sizes (between 250, 500, and 1000) to compare different IRT models. One study extended stochastic curtailment to polytomous models with multiple cut-points (Gnambs and Batinic, 2011), but the authors of that study did not provide the equations that they used, and their stopping rules were similarly arbitrary extensions of Finkelman (2008a) to more than two categories. Additional comparisons of CCT termination criteria among the profusion of specialty IRT models (including polytomous, multidimensional, cognitive diagnoses models, etc.) is not sensible at present.

As a practical concern, item banks for a future study should be based off of realistic exams. Bartroff et al. (2008) and Finkelman (2008a) used 1,136 items from a subsidiary of the Educational Testing Service; Gnambs and Batinic (2011) used various numbers of items from traditional personality measures; and Wouda and Eggen (2009) used 250 items from a mathematics item bank. Thompson (2009, 2010) always simulated item banks to retain certain distributional properties, which is unwise considering that all methods must eventually be applied to serviceable exams. In the current study, 600 items were used from a real item bank, yet those items were only a selection from the full bank. As an extension of this study, an additional 400 items will be appropriated, and then subsets of the full 1000 item bank will be randomly selected to form the smaller item pools. Because all items were calibrated under the 3PL, 1000 items will be generated under the 2PL to have similar test bank properties.

Even after varying termination criteria and choosing an item bank, ability estimation and item selection algorithms are important for the performance of adaptive tests. In the current study, varying the ability estimation and item selection algorithm did not seem to crucially affect classification accuracy and test length. Yet ability estimation is an integral part of stochastic curtailment, and weighted likelihood estimation did appear to slightly improve classification accuracy for SCSPRT conditions over MLE (see Figure D.4 in Appendix D), even though the gain in accuracy might have been hidden by other, more relevant factors, such as exposure control. Forming a confidence interval around $\hat{\theta}$ before testing classification probabilities is arbitrary, and other estimation procedures, such as weighted likelihood estimation, Bayesian modal estimation (BME), and expected a posteriori estimation (EAP) overcome the outward bias of the latent trait estimate without relying as much on item properties (e.g., van der Linden, 2010). Item selection is also necessary for determining the key probabilities that stochastic curtailment uses to make a decision. In addition, choosing items appropriately can drastically reduce expected test length. As was illustrated in Section 4.3, the log-likelihood ratio test statistic used in the SPRT behaves counterintuitively in the 3PL for examinees with true ability much higher than a classification boundary. One method of improving the performance of simple likelihood ratio tests is to select more items at the closest cutpoint, but additional work is needed to determine the best method for selecting items when there are multiple cutting points.

The simulations presented herein demonstrate the power of likelihood ratio-based methods for efficiently and accurately classifying examinees when there are multiple categories. The proposed study will extend previous work to better estimate classification probabilities, better control error rates, and better determine exam properties that result in the most accurate classifications using the least number of items. Adaptive testing is already being touted as the future of high stakes exams, but only when practitioners are knowledgeable of the ideal testing procedure 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

ACT. (2007). *WorkKeys: An Overview* [Brochure]. Retrieved November 14, 2011, from: http://www.edgecombe.edu/crc/pdfs/workkeys_overview.pdf

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 multiple hypotheses. *Communications in Statistics – Theory and Methods*, *39*, 1597–1607.

Bejar, I. I. (1983). *Achievement testing: Recent advances.* Beverly Hills, CA: Sage Publications.

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.

Blinnikov, S., & Moessner, R. (1998). Expansions for nearly Gaussian distributions. *Astronomy & Astrophysics Supplement Series*, *130*, 193–205.

Camilli, G. (1994). Origin of the scaling constant d = 1.7 in item response theory. *Journal of Educational and Behavioral Statistics*, *19*, 293–296.

Casella, G., & Berger, R. L. (2001). *Statistical inference*, Pacific Grove, CA: Duxbury Press.

Common Core State Standards Initiative. (2010). *Common Core State Standards*. Retrieved from http://www.corestandards.org/

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., Ankenmann, R. D., & Chang, H.-H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, *24*, 241–255.

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

DiBello, L. V., & Stout, W. (Eds.). (2007). IRT-based cognitive diagnostic models and related methods [Special issues]. *Journal of Educational Measurement*, *44*.

Dmitrienko, A. & Wang, M.-D. (2006). Bayesian predictive approach to interim monitoring in clinical trials. *Statistics in Medicine*, *25*, 2178–2195.

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.

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

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

FInkelman, M. (2008b). The Wald-Wolfowitz theorem is violated in sequential mastery testing. *Sequential Analysis*, *27*, 293–303.

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

Finkelman, M. D., Hooker, G., & Wang, Z. (2010). Prevalence and magnitude of paradoxical results in multidimensional item response theory. *Journal of Educational and Behavioral Statistics*, *35*, 744–761.

Frank, S. A. (2009). Natural selection maximizes Fisher information. *Journal of Evolutionary Biology*, *22*, 231–244.

Gnambs, T., & Batinic, B. (2011). Polytomous adaptive classification testing: Effects of item pool size, test termination criterion, and number of cutscores. *Educational and Psychological Measurement*, *71*, 1006–1022.

Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with graduate process change. *Psychometrika*, *73*, 65–87.

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

Hooker, G. (2010). On separable tests, correlated priors, and paradoxical results in multidimensional item response theory. *Psychometrika*, *75*, 694–707.

Hooker, G., & Finkelman, M. (2010). Paradoxical results and item bundles. *Psychometrika*, *75*, 249–271.

Hooker, G., Finkelman, M., & Schwartzman, A. (2009). Paradoxical results in multidimensional item response theory. *Psychometrika*, *74*, 419–442.

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Boca Raton, FL: Chapman & Hall.

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.

Lai, T. L. (1997). On optimal stopping problems in sequential hypothesis testing. *Statistical Sinica*, *7*, 33–51.

Lai, T. L. (2001). Sequential analysis: Some classical problems and new challenges. *Statistical Sinica*, *11*, 303–408.

Lai, T. L., & Shih, M. C. (2004). Power, sample size, and adaptation considerations in the design of group sequential clinical trials. *Biometrika*, *91*, 507–528.

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.

Mislevy, R. J., & Chang, H.-H. (2000). Does adaptive testing violate local independence. *Psychometrika*, *65*, 149–156.

No Child Left Behind Act of 2001, 20 U.S.C. §6319 (2008).

Nydick, S. W., Nozawa, Y., & Zhu, R. (2012, April). *Accuracy and efficiency in classifying examinees using computerized adaptive tests: An application to a large scale test.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*, 241–286.

Paulson, E. (1963). A sequential decision procedure for choosing between one of k hypotheses concerning the unknown mean of a normal distribution. *Annals of Mathematical Statistics*, *34*, 549–554.

Pavlov, I. V. (1988). A sequential procedure for testing many composite hypotheses. *Theory of Probability and its Applications*, *33*, 138–142.

Pearson. (2011). *New York State Teacher Certification Exams*. Retrieved November 14, 2011, from: http://www.nystce.nesinc.com/index.asp

R Development Core Team (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: http://www.R-project.org/.

Reckase, M. D. (1979, June). *Some decision procedures for use with tailored testing*. Paper presented at the Meeting of Computer Assisted Testing, Minneapolis Minnesota.

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

Revuelta, J. (2008). The generalized logit-linear item response model for binary-designed items. *Psychometrika*, *73*, 385–405.

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, LA.

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. (2009). 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, NL.

Vos, H. J., & Glas, C. A. W. (2010). Testlet-based adaptive mastery testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing*. New York, NY: Springer.

van der Linden, W. J. (2005). A comparison of item selection methods for adaptive tests with content constraints. *Journal of Educational Measurement*, *42*, 3, 283–302.

van der Linden, W. J. & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing*. New York, NY: Springer.

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

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, *16*, 117–186.

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

Wald, A., & Wolfowitz, J. (1948). Optimal character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, *19*, 326–339

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.

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

Weng, R. C. (2010). A Bayesian Edgeworth expansion by Stein's identity. *Bayesian Analysis*, *5*, 741–764.

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. *Educational and Psychological Measurement*, *66*, 545–564.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (2006). *BILOG-MG 3 for Windows*. Chicago, IL: Scientific Software International.

Appendix A

Tables: Conditional on Ability

The following tables indicate the conditional accuracy and test length for various conditions at θ_i near each of the cut-points. In all of the tables, a Sympson-Hetter item exposure control method was implemented, as described in Section 3.4.

Table A.1: Test length averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at the nearest cut-point, ability estimated by maximum likelihood estimation, and no item exposure control. The midpoint between each of the vertical bars is close to a classification bound, and only a few θ_i on either side of each classification bound are presented for clarity.

Condition	$\theta_i = -1.6$	-1.5	-1.4	-1.3	-1.2	-0.7	-0.6	-0.5	-0.4	-0.3
SPRT: $\delta = .10$	20.9	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .15$	16.7	20.9	11.9	20.9	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .20$	18.4	19.9	20.1	20.5	20.6	20.7	20.8	20.9	20.8	20.8
SPRT: $\delta = .25$	17.8	19.0	19.4	18.5	19.6	18.9	19.4	19.7	20.1	19.9
SPRT: $\delta = .30$	16.0	16.6	17.7	18.8	18.8	18.5	17.3	18.8	18.9	18.8
SCSPRT: $\delta = .20$; $\gamma = .95$	17.0	17.7	17.7	17.7	18.4	18.4	18.8	18.5	19.0	19.0
SCSPRT: $\delta = .20$; $\gamma = .86$	15.9	16.8	17.1	17.4	17.9	18.2	18.6	18.6	18.9	18.7
SCSPRT: $\delta = .20$; $\gamma = .68$	16.2	16.9	17.4	17.6	17.4	17.5	18.2	18.3	18.3	18.3
SCSPRT: $\delta = .20$; $\gamma = .38$	16.0	16.3	17.3	18.0	16.5	17.2	17.7	18.1	17.6	18.3
SCSPRT: $\delta = .20$; $\gamma = .00$	13.4	14.1	16.0	16.0	16.7	16.6	17.7	17.1	17.4	17.0
GLR: $\delta = .10$	18.8	19.5	19.5	20.3	20.7	21.0	21.0	21.0	21.0	21.0
GLR: $\delta = .15$	16.2	18.4	19.8	20.8	19.2	20.3	20.9	20.9	20.9	21.0
GLR: $\delta = .20$	17.4	18.7	19.5	19.5	19.9	20.2	20.5	20.4	20.8	20.7
GLR: $\delta = .25$	16.9	17.0	18.7	19.0	19.1	19.4	19.2	19.8	20.2	19.6
GLR: $\delta = .30$	15.0	16.6	17.0	17.9	18.0	17.7	18.4	18.4	18.4	18.7

Condition	$\theta_i = 0.1$	0.2	0.3	0.4	0.5	1.0	1.1	1.2	1.3	1.4
SPRT: $\delta = .10$	21.0	21.0	21.0	21.0	21.0	21.0	20.9	20.9	20.9	20.9
SPRT: $\delta = .15$	21.0	21.0	21.0	21.0	20.9	20.9	21.0	20.9	20.9	15.0
SPRT: $\delta = .20$	20.8	20.6	20.6	20.1	19.7	19.3	19.4	18.5	18.3	17.0
SPRT: $\delta = .25$	19.9	19.9	18.9	17.7	17.6	16.6	17.1	17.3	17.0	15.5
SPRT: $\delta = .30$	17.7	17.8	17.7	16.4	16.4	14.6	15.2	15.8	15.6	13.6
SCSPRT: $\delta = .20$; $\gamma = .95$	19.0	19.0	18.1	18.1	18.2	17.7	17.1	18.1	17.2	16.4
SCSPRT: $\delta = .20$; $\gamma = .86$	18.6	18.7	18.5	17.8	17.7	16.9	17.1	17.1	16.7	15.3
SCSPRT: $\delta = .20$; $\gamma = .68$	18.4	18.0	18.0	17.4	17.1	16.8	16.2	16.9	16.7	15.4
SCSPRT: $\delta = .20$; $\gamma = .38$	18.3	17.5	17.5	17.1	16.0	16.3	16.2	15.5	15.3	13.1
SCSPRT: $\delta = .20$; $\gamma = .00$	16.5	16.2	17.0	16.7	15.5	14.5	16.3	15.3	15.1	13.5
GLR: $\delta = .10$	21.0	20.8	20.8	20.5	20.3	19.7	19.8	19.1	18.8	17.7
GLR: $\delta = .15$	21.0	19.9	20.7	19.2	19.3	20.4	20.0	18.3	17.9	16.2
GLR: $\delta = .20$	20.2	19.9	20.2	19.2	18.3	17.3	17.7	18.1	15.9	15.0
GLR: $\delta = .25$	19.6	19.1	18.6	17.0	17.3	15.3	16.5	17.0	15.7	13.0
GLR: $\delta = .30$	17.8	17.3	17.1	16.7	16.1	14.8	15.1	14.5	14.4	13.0

Table A.2: Classification accuracy averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at the nearest cut-point, ability estimated by maximum likelihood estimation, and no item exposure control. The midpoint between each of the vertical bars is close to a classification bound, and only a few θ_i on either side of each classification bound are presented for clarity.

Condition	$\theta_i = -1.6$	-1.5	-1.4	-1.3	-1.2	-0.7	-0.6	-0.5	-0.4	-0.3
SPRT: $\delta = .10$	.77	.66	.52	.73	.77	.79	.64	.56	.63	.74
SPRT: $\delta = .15$	.96	.11	.93	.94	.95	.98	.12	.92	.93	.96
SPRT: $\delta = .20$	.82	.63	.46	.70	.83	.75	.62	.48	.60	.73
SPRT: $\delta = .25$	.74	.52	.44	.60	.74	.78	.68	.58	.61	.71
SPRT: $\delta = .30$	.76	.70	.51	.56	.77	.86	.74	.55	.53	.76
SCSPRT: $\delta = .20$; $\gamma = .95$	.81	.62	.40	.58	.67	.72	.56	.58	.70	.75
SCSPRT: $\delta = .20$; $\gamma = .86$	.74	.66	.54	.61	.72	.78	.64	.46	.64	.78
SCSPRT: $\delta = .20$; $\gamma = .68$	.79	.68	.55	.66	.77	.79	.64	.42	.58	.74
SCSPRT: $\delta = .20$; $\gamma = .38$	.79	.57	.46	.67	.66	.76	.69	.54	.55	.64
SCSPRT: $\delta = .20$; $\gamma = .00$	.78	.70	.57	.55	.75	.69	.60	.53	.65	.72
GLR: $\delta = .10$	.77	.57	.58	.65	.74	.71	.74	.52	.73	.71
GLR: $\delta = .15$	.77	.75	.36	.60	.64	.70	.64	.58	.65	.71
GLR: $\delta = .20$	.81	.61	.49	.56	.82	.83	.62	.51	.57	.77
GLR: $\delta = .25$	.82	.63	.52	.64	.80	.87	.62	.53	.68	.80
GLR: $\delta = .30$	.82	.64	.52	.62	.69	.76	.66	.54	.61	.75

Condition	$\theta_i = 0.1$	0.2	0.3	0.4	0.5	1.0	1.1	1.2	1.3	1.4
SPRT: $\delta = .10$	.70	.62	.54	.71	.86	.77	.63	.50	.74	.82
SPRT: $\delta = .15$	.14	.92	.94	.96	.97	.96	.95	.90	.94	.97
SPRT: $\delta = .20$	.77	.65	.52	.76	.83	.80	.65	.54	.76	.89
SPRT: $\delta = .25$	.79	.68	.54	.74	.91	.66	.69	.54	.71	.9
SPRT: $\delta = .30$	.78	.67	.57	.58	.78	.76	.52	.58	.64	.9
SCSPRT: $\delta = .20$; $\gamma = .95$	.81	.62	.60	.78	.75	.64	.63	.47	.74	.85
SCSPRT: $\delta = .20$; $\gamma = .86$	.68	.59	.55	.71	.80	.76	.60	.50	.68	.84
SCSPRT: $\delta = .20$; $\gamma = .68$	.73	.58	.58	.72	.84	.80	.58	.49	.68	.86
SCSPRT: $\delta = .20$; $\gamma = .38$	.66	.61	.62	.66	.85	.86	.58	.50	.68	.87
SCSPRT: $\delta = .20$; $\gamma = .00$	.63	.55	.53	.74	.80	.71	.68	.65	.62	.81
GLR: $\delta = .10$	.65	.57	.56	.75	.83	.80	.63	.62	.76	.89
GLR: $\delta = .15$	.76	.54	.53	.65	.88	.83	.42	.48	.76	.88
GLR: $\delta = .20$	.80	.56	.46	.65	.85	.75	.59	.58	.73	.78
GLR: $\delta = .25$	.76	.71	.64	.75	.86	.74	.66	.55	.81	.86
GLR: $\delta = .30$	.74	.61	.56	.68	.82	.76	.65	.54	.72	.88

Table A.3: Test length averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at $\hat{\theta}_i$, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\text{max}} = .2$. The midpoint between each of the vertical bars is close to a classification bound, and only a few θ_i on either side of each classification bound are presented for clarity.

Condition	$\theta_i = -1.6$	-1.5	-1.4	-1.3	-1.2	-0.7	-0.6	-0.5	-0.4	-0.3
SPRT: $\delta = .10$	20.9	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .15$	20.7	20.6	20.9	21.0	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .20$	19.6	20.2	20.1	20.5	20.4	20.5	20.7	20.5	20.9	20.9
SPRT: $\delta = .25$	18.1	18.8	19.6	19.5	19.3	19.8	20.0	20.1	20.1	20.2
SPRT: $\delta = .30$	16.1	17.5	17.5	19.0	18.4	18.4	18.9	18.8	19.4	19.0
SCSPRT: $\delta = .20$; $\gamma = .95$	16.5	17.9	18.7	18.7	18.4	18.7	18.9	18.8	19.2	19.3
SCSPRT: $\delta = .20$; $\gamma = .86$	16.7	18.0	18.4	17.6	18.3	17.3	18.0	19.1	18.4	19.4
SCSPRT: $\delta = .20$; $\gamma = .68$	16.3	19.0	18.7	14.0	17.3	14.7	18.3	18.0	19.3	20.7
SCSPRT: $\delta = .20$; $\gamma = .38$	16.6	16.7	16.5	16.7	16.9	17.1	17.1	17.9	17.1	18.0
SCSPRT: $\delta = .20$; $\gamma = .00$	14.0	15.6	15.8	16.2	16.2	17.0	17.0	17.3	17.3	16.8
GLR: $\delta = .10$	19.3	17.7	19.3	19.7	20.9	21.0	20.9	21.0	21.0	21.0
GLR: $\delta = .15$	17.8	18.4	20.0	19.7	20.6	20.6	20.9	20.7	21.0	20.9
GLR: $\delta = .20$	17.1	18.4	20.2	18.5	19.9	20.7	20.8	20.6	20.9	20.9
GLR: $\delta = .25$	16.4	17.7	17.9	17.9	19.0	18.5	19.8	19.9	19.7	19.7
GLR: $\delta = .30$	15.8	17.0	17.6	17.9	18.8	17.9	18.5	19.5	19.2	19.1

Condition	$\theta_i = 0.1$	0.2	0.3	0.4	0.5	1.0	1.1	1.2	1.3	1.4
SPRT: $\delta = .10$	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .15$	21.0	20.9	20.9	20.9	20.8	20.6	20.9	20.8	20.7	20.5
SPRT: $\delta = .20$	20.7	20.8	20.3	20.4	19.4	19.4	20.0	19.6	19.3	18.3
SPRT: $\delta = .25$	19.8	19.4	19.3	19.1	17.6	17.8	17.8	18.0	18.1	16.8
SPRT: $\delta = .30$	18.6	18.9	18.0	17.7	16.5	15.8	17.0	16.7	15.7	15.6
SCSPRT: $\delta = .20$; $\gamma = .95$	18.7	19.2	19.0	18.3	17.9	17.3	17.8	17.9	17.5	16.8
SCSPRT: $\delta = .20$; $\gamma = .86$	18.6	19.2	18.2	17.9	18.8	17.5	18.1	16.4	17.4	14.5
SCSPRT: $\delta = .20$; $\gamma = .68$	19.3	18.7	16.7	20.7	18.7	15.7	16.0	16.0	20.3	15.0
SCSPRT: $\delta = .20$; $\gamma = .38$	17.8	18.2	18.0	16.9	17.2	16.8	16.3	15.5	15.7	14.8
SCSPRT: $\delta = .20$; $\gamma = .00$	16.9	16.8	16.3	16.0	15.1	15.7	16.0	15.7	14.9	14.7
GLR: $\delta = .10$	21.0	20.9	20.9	20.4	20.5	19.5	19.2	19.0	19.5	16.9
GLR: $\delta = .15$	20.8	20.4	20.6	19.8	19.2	19.0	19.3	18.4	18.1	17.1
GLR: $\delta = .20$	20.5	20.6	20.6	18.9	19.9	19.9	18.0	16.4	19.6	18.7
GLR: $\delta = .25$	19.7	19.6	18.6	17.3	17.9	17.0	17.4	16.6	16.4	15.4
GLR: $\delta = .30$	17.9	19.0	17.4	16.0	17.2	15.1	14.6	15.8	15.7	13.4

Table A.4: Classification accuracy averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at $\hat{\theta}_i$, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\max} = .2$. The midpoint between each of the vertical bars is close to a classification bound, and only a few θ_i on either side of each classification bound are presented for clarity.

Condition	$\theta_i = -1.6$	-1.5	-1.4	-1.3	-1.2	-0.7	-0.6	-0.5	-0.4	-0.3
SPRT: $\delta = .10$	.77	.70	.58	.66	.74	.84	.70	.52	.57	.71
SPRT: $\delta = .15$	.71	.61	.52	.73	.76	.83	.63	.40	.64	.78
SPRT: $\delta = .20$	.82	.73	.39	.67	.74	.67	.69	.63	.77	.70
SPRT: $\delta = .25$	.77	.61	.49	.65	.73	.81	.62	.52	.60	.66
SPRT: $\delta = .30$	.76	.70	.52	.62	.71	.81	.68	.52	.59	.70
SCSPRT: $\delta = .20; \gamma = .95$	.73	.70	.52	.67	.78	.75	.65	.48	.63	.73
SCSPRT: $\delta = .20; \gamma = .86$	.77	.70	.41	.37	.82	.82	.54	.53	.65	.60
SCSPRT: $\delta = .20; \gamma = .68$	.67	.33	.00	.33	1.00	.34	1.00	1.00	.33	.67
SCSPRT: $\delta = .20; \gamma = .38$	.82	.66	.48	.65	.76	.74	.62	.49	.70	.70
SCSPRT: $\delta = .20; \gamma = .00$	.79	.66	.52	.60	.74	.74	.58	.48	.66	.76
GLR: $\delta = .10$	.85	.64	.48	.64	.72	.75	.74	.52	.57	.76
GLR: $\delta = .15$	.70	.65	.49	.53	.83	.79	.60	.58	.62	.68
GLR: $\delta = .20$	.72	.88	.42	.66	.70	.73	.89	.64	.63	.46
GLR: $\delta = .25$	.77	.70	.54	.52	.75	.78	.64	.50	.56	.77
GLR: $\delta = .30$	.81	.62	.53	.54	.71	.71	.67	.55	.71	.78

Condition	$\theta_i = 0.1$	0.2	0.3	0.4	0.5	1.0	1.1	1.2	1.3	1.4
SPRT: $\delta = .10$	.72	.60	.53	.71	.81	.74	.63	.55	.76	.85
SPRT: $\delta = .15$	.68	.61	.50	.71	.80	.75	.63	.59	.70	.89
SPRT: $\delta = .20$	.70	.70	.46	.54	.74	.86	.60	.64	.76	.86
SPRT: $\delta = .25$	.74	.62	.52	.66	.88	.76	.58	.57	.73	.89
SPRT: $\delta = .30$	.77	.69	.58	.72	.75	.82	.60	.46	.68	.87
SCSPRT: $\delta = .20; \gamma = .95$	.66	.60	.47	.74	.83	.67	.62	.56	.69	.81
SCSPRT: $\delta = .20; \gamma = .86$	.82	.59	.42	.70	.82	.88	.53	.65	.87	.88
SCSPRT: $\delta = .20; \gamma = .68$	1.00	.67	.67	1.00	.33	1.00	.67	.34	1.00	1.00
SCSPRT: $\delta = .20; \gamma = .38$	.71	.59	.60	.74	.74	.64	.68	.69	.76	.93
SCSPRT: $\delta = .20; \gamma = .00$	.70	.56	.53	.72	.84	.71	.58	.53	.78	.85
GLR: $\delta = .10$	.76	.63	.62	.76	.73	.82	.52	.62	.66	.89
GLR: $\delta = .15$	.82	.51	.42	.65	.77	.66	.68	.63	.54	.92
GLR: $\delta = .20$	.92	.43	.41	.90	.72	.70	.64	.86	.92	.73
GLR: $\delta = .25$	.70	.58	.72	.64	.73	.77	.62	.60	.78	.86
GLR: $\delta = .30$	.68	.65	.62	.72	.82	.76	.42	.52	.72	.8

Table A.5: Test length averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at the nearest cut-point, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\max} = .2$. The midpoint between each of the vertical bars is close to a classification bound, and only a few θ_i on either side of each classification bound are presented for clarity.

Condition	$\theta_i = -1.6$	-1.5	-1.4	-1.3	-1.2	-0.7	-0.6	-0.5	-0.4	-0.3
SPRT: $\delta = .10$	20.9	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .15$	19.5	20.9	20.9	21.0	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .20$	19.0	20.0	20.0	20.3	20.5	20.9	21.0	20.7	21.0	20.8
SPRT: $\delta = .25$	18.6	18.4	19.9	19.6	19.0	20.0	20.2	20.2	20.1	20.3
SPRT: $\delta = .30$	18.6	16.6	16.3	18.0	16.7	19.2	19.3	18.8	19.0	18.6
SCSPRT: $\delta = .20; \gamma = .95$	16.7	18.3	16.7	18.1	19.3	18.7	19.0	18.9	18.9	19.4
SCSPRT: $\delta = .20; \gamma = .86$	16.5	16.7	17.0	18.5	17.6	18.5	19.0	18.4	18.9	18.7
SCSPRT: $\delta = .20; \gamma = .68$	15.7	17.1	17.8	17.7	18.0	18.4	18.2	17.2	18.9	19.4
SCSPRT: $\delta = .20; \gamma = .38$	16.4	15.4	17.4	17.1	16.8	17.6	17.7	17.7	18.4	18.0
SCSPRT: $\delta = .20; \gamma = .00$	13.5	16.6	16.2	15.4	16.6	16.8	17.7	17.7	17.9	17.3
GLR: $\delta = .10$	18.1	18.7	20.3	20.4	20.0	21.0	21.0	21.0	21.0	21.0
GLR: $\delta = .15$	17.7	19.4	19.1	19.2	20.5	20.8	21.0	20.9	21.0	20.9
GLR: $\delta = .20$	17.4	17.4	19.2	19.5	19.9	20.4	20.5	20.6	20.9	20.7
GLR: $\delta = .25$	15.8	17.1	18.5	18.9	18.9	19.0	19.8	19.9	20.1	20.0
GLR: $\delta = .30$	15.3	16.0	17.0	17.7	18.2	17.9	18.7	18.9	18.9	18.8

Condition	$\theta_i = 0.1$	0.2	0.3	0.4	0.5	1.0	1.1	1.2	1.3	1.4
SPRT: $\delta = .10$	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	20.8
SPRT: $\delta = .15$	21.0	21.0	21.0	20.9	21.0	21.0	20.9	20.1	19.4	18.9
SPRT: $\delta = .20$	20.8	20.8	20.6	20.1	19.9	20.1	20.1	19.0	19.1	17.1
SPRT: $\delta = .25$	20.1	19.7	19.6	19.1	18.5	17.9	18.1	16.3	15.8	15.6
SPRT: $\delta = .30$	18.9	19.4	17.0	17.4	19.0	13.7	16.4	15.3	15.9	11.4
SCSPRT: $\delta = .20; \gamma = .95$	19.1	18.8	19.0	18.3	17.4	17.0	17.0	16.5	18.2	15.9
SCSPRT: $\delta = .20; \gamma = .86$	19.1	18.6	18.6	18.8	17.6	16.0	16.2	17.8	16.9	14.5
SCSPRT: $\delta = .20; \gamma = .68$	18.6	17.3	18.5	18.0	17.3	15.5	16.5	17.4	15.9	14.3
SCSPRT: $\delta = .20; \gamma = .38$	17.5	17.6	16.9	17.1	16.7	16.5	17.1	16.9	14.8	14.8
SCSPRT: $\delta = .20; \gamma = .00$	16.9	16.6	17.8	15.5	15.4	15.2	15.2	15.9	13.7	13.8
GLR: $\delta = .10$	20.9	20.8	20.8	20.5	20.4	19.7	19.9	19.4	18.6	17.3
GLR: $\delta = .15$	20.6	20.9	20.6	20.4	18.5	19.5	17.7	18.7	18.1	15.4
GLR: $\delta = .20$	20.4	19.9	20.1	19.3	18.3	16.9	17.9	17.9	16.8	16.2
GLR: $\delta = .25$	19.5	19.3	18.3	18.4	17.1	16.3	17.7	15.8	15.6	15.6
GLR: $\delta = .30$	18.1	18.3	17.8	17.1	15.9	15.5	15.3	14.4	14.1	14.3

Table A.6: Classification accuracy averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at the nearest cut-point, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\max} = .2$. The midpoint between each of the vertical bars is close to a classification bound, and only a few θ_i on either side of each classification bound are presented for clarity.

Condition	$\theta_i = -1.6$	-1.5	-1.4	-1.3	-1.2	-0.7	-0.6	-0.5	-0.4	-0.3
SPRT: $\delta = .10$	.81	.63	.46	.68	.76	.75	.67	.54	.65	.77
SPRT: $\delta = .15$	.64	.79	.77	.76	.79	.65	.66	.38	.51	.70
SPRT: $\delta = .20$	.75	.62	.50	.52	.72	.86	.71	.62	.65	.78
SPRT: $\delta = .25$	.76	.62	.33	.66	.72	.77	.64	.60	.56	.72
SPRT: $\delta = .30$	.98	.84	.31	.70	.72	.60	.58	.44	.69	.84
SCSPRT: $\delta = .20; \gamma = .95$	.75	.45	.62	.64	.78	.79	.72	.47	.44	.77
SCSPRT: $\delta = .20; \gamma = .86$	.63	.63	.67	.73	.70	.69	.69	.39	.68	.74
SCSPRT: $\delta = .20; \gamma = .68$	.99	.45	.66	.66	.78	.67		.34	.44	.34
SCSPRT: $\delta = .20; \gamma = .38$	.75	.67	.67	.61	.68	.80	.68	.52	.70	.78
SCSPRT: $\delta = .20; \gamma = .00$	.84	.46	.44	.40	.67	.75	.50	.51	.49	.64
GLR: $\delta = .10$	.85	.73	.49	.63	.74	.78	.63	.48	.57	.81
GLR: $\delta = .15$	.68	.71	.49	.60	.82	.78	.68	.64	.64	.64
GLR: $\delta = .20$	.83	.70	.37	.57	.81	.70	.68	.34	.68	.77
GLR: $\delta = .25$	.77	.66	.50	.66	.80	.66	.81	.47	.57	.77
GLR: $\delta = .30$	.78	.64	.43	.62	.72	.81	.67	.63	.55	.74

Condition	$\theta_i = 0.1$	0.2	0.3	0.4	0.5	1.0	1.1	1.2	1.3	1.4
SPRT: $\delta = .10$	.72	.64	.64	.70	.86	.76	.63	.55	.69	.82
SPRT: $\delta = .15$	.85	.63	.43	.83	.98	.85	.78	.83	.78	.98
SPRT: $\delta = .20$	.76	.69	.58	.72	.87	.72	.66	.56	.75	.87
SPRT: $\delta = .25$	.67	.62	.51	.65	.80	.84	.61	.50	.78	.84
SPRT: $\delta = .30$	.72	.83	.56	.70	.35	.84	.82	.32	.46	.85
SCSPRT: $\delta = .20; \gamma = .95$	.64	.62	.48	.60	.79	.83	.72	.51	.64	.83
SCSPRT: $\delta = .20; \gamma = .86$	.63	.67	.45	.40	.81	.98	.63	.27	.80	.76
SCSPRT: $\delta = .20; \gamma = .68$	.99	.78	.78	.88	.78	.89	.56	.34	.56	.89
SCSPRT: $\delta = .20; \gamma = .38$	.79	.73	.56	.66	.86	.88	.70	.51	.82	.82
SCSPRT: $\delta = .20; \gamma = .00$	.76	.51	.50	.71	.81	.81	.62	.48	.68	.85
GLR: $\delta = .10$	.81	.58	.54	.69	.84	.78	.60	.62	.73	.85
GLR: $\delta = .15$	.77	.71	.66	.57	.93	.81	.60	.50	.53	.96
GLR: $\delta = .20$	.78	.62	.51	.66	.86	.76	.57	.55	.81	.76
GLR: $\delta = .25$	.77	.62	.70	.71	.85	.75	.64	.50	.76	.83
GLR: $\delta = .30$	.76	.56	.52	.62	.79	.74	.62	.54	.75	.84

Table A.7: Test length averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at $\hat{\theta}_i$, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\max} = .1$. The midpoint between each of the vertical bars is close to a classification bound, and only a few θ_i on either side of each classification bound are presented for clarity.

Condition	$\theta_i = -1.6$	-1.5	-1.4	-1.3	-1.2	-0.7	-0.6	-0.5	-0.4	-0.3
SPRT: $\delta = .10$	21.0	21.0	21.0	20.9	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .15$	20.3	20.8	20.9	21.0	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .20$	18.8	19.7	20.3	20.8	20.6	21.0	21.0	20.9	21.0	21.0
SPRT: $\delta = .25$	16.6	18.8	20.3	20.4	20.4	20.6	20.5	20.8	20.9	20.8
SPRT: $\delta = .30$	17.1	18.7	18.6	18.8	19.0	19.3	19.8	20.2	20.1	19.4
SCSPRT: $\delta = .20$; $\gamma = .95$	17.4	15.8	16.8	18.8	18.4	18.5	19.0	18.4	19.4	19.0
SCSPRT: $\delta = .20$; $\gamma = .86$	17.1	17.6	18.6	18.3	18.3	18.7	18.8	19.2	19.0	19.1
SCSPRT: $\delta = .20$; $\gamma = .68$	16.8	17.3	17.6	18.4	18.5	18.3	18.5	18.9	18.9	18.9
SCSPRT: $\delta = .20$; $\gamma = .38$	16.6	17.1	17.0	17.5	17.6	18.1	18.3	18.4	18.3	18.4
SCSPRT: $\delta = .20$; $\gamma = .00$	17.1	16.3	16.0	18.1	17.0	16.6	17.9	17.8	17.9	17.8
GLR: $\delta = .10$	18.5	19.3	19.4	20.3	20.2	21.0	21.0	21.0	21.0	20.9
GLR: $\delta = .15$	17.7	18.5	19.2	19.5	20.4	20.9	20.8	21.0	21.0	21.0
GLR: $\delta = .25$	17.1	17.9	19.1	18.9	19.7	20.0	20.3	20.5	20.5	20.3
GLR: $\delta = .30$	15.4	17.2	18.4	18.4	18.7	19.2	19.3	19.4	20.0	19.8

Condition	$\theta_i = 0.1$	0.2	0.3	0.4	0.5	1.0	1.1	1.2	1.3	1.4
SPRT: $\delta = .10$	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .15$	21.0	21.0	21.0	21.0	21.0	20.9	20.9	20.8	20.8	20.8
SPRT: $\delta = .20$	20.9	20.9	20.6	20.5	20.2	19.8	20.2	20.0	20.1	19.5
SPRT: $\delta = .25$	21.0	19.8	19.7	19.6	19.6	17.6	20.0	19.4	18.4	18.3
SPRT: $\delta = .30$	19.2	19.6	19.5	18.6	17.7	17.0	16.6	16.8	15.8	16.7
SCSPRT: $\delta = .20$; $\gamma = .95$	19.6	18.8	20.5	19.3	18.4	17.5	17.6	18.0	16.7	17.5
SCSPRT: $\delta = .20$; $\gamma = .86$	19.5	18.9	18.2	18.7	18.2	16.6	16.5	17.0	17.3	15.6
SCSPRT: $\delta = .20$; $\gamma = .68$	18.9	18.6	18.4	18.4	17.8	17.2	17.1	16.6	15.7	15.0
SCSPRT: $\delta = .20$; $\gamma = .38$	18.1	18.0	18.3	17.4	16.8	16.3	16.9	16.1	15.1	14.9
SCSPRT: $\delta = .20$; $\gamma = .00$	17.7	16.8	16.7	16.8	15.3	15.6	16.1	16.5	15.4	14.4
GLR: $\delta = .10$	21.0	21.0	20.9	20.8	20.6	20.2	19.8	18.5	19.2	17.7
GLR: $\delta = .15$	20.9	20.8	20.6	20.3	20.2	20.0	19.6	19.0	18.0	17.0
GLR: $\delta = .20$	21.0	20.7	20.0	20.7	19.5	18.5	18.0	18.3	17.5	17.9
GLR: $\delta = .25$	19.8	19.8	19.4	19.2	18.5	17.6	17.6	17.6	15.6	14.5
GLR: $\delta = .30$	19.1	18.8	18.4	18.5	17.3	16.6	16.2	16.5	15.6	14.4

Table A.8: Classification accuracy averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at $\hat{\theta}_i$, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\max} = .1$. The midpoint between each of the vertical bars is close to a classification bound, and only a few θ_i on either side of each classification bound are presented for clarity.

Condition	$\theta_i = -1.6$	-1.5	-1.4	-1.3	-1.2	-0.7	-0.6	-0.5	-0.4	-0.3
SPRT: $\delta = .10$	.76	.56	.38	.62	.65	.64	.64	.58	.64	.78
SPRT: $\delta = .15$	.82	.70	.52	.62	.75	.78	.61	.54	.61	.64
SPRT: $\delta = .20$	.78	.65	.50	.62	.75	.78	.69	.49	.67	.78
SPRT: $\delta = .25$	.78	.71	.22	.78	.57	.64	.57	.64	.65	.72
SPRT: $\delta = .30$	.66	.71	.54	.75	.76	.72	.68	.49	.55	.68
SCSPRT: $\delta = .20; \gamma = .95$	.81	.61	.59	.59	.98	.98	.60	.78	.77	.61
SCSPRT: $\delta = .20; \gamma = .86$	.73	.62	.48	.62	.68	.75	.60	.60	.54	.67
SCSPRT: $\delta = .20; \gamma = .68$	.72	.64	.54	.61	.67	.72	.64	.53	.62	.76
SCSPRT: $\delta = .20; \gamma = .38$	.66	.60	.58	.57	.63	.79	.56	.48	.58	.77
SCSPRT: $\delta = .20; \gamma = .00$	.62	.63	.43	.70	.62	.72	.70	.44	.47	.64
GLR: $\delta = .10$	.80	.63	.49	.59	.65	.83	.63	.56	.59	.75
GLR: $\delta = .15$	.77	.64	.50	.63	.74	.80	.68	.54	.60	.70
GLR: $\delta = .20$	.84	.55	.54	.84	.98	.63	.77	.55	.40	.83
GLR: $\delta = .25$	.78	.66	.48	.62	.65	.80	.66	.56	.51	.76
GLR: $\delta = .30$	.78	.64	.53	.61	.71	.72	.70	.61	.63	.74

Condition	$\theta_i = 0.1$	0.2	0.3	0.4	0.5	1.0	1.1	1.2	1.3	1.4
SPRT: $\delta = .10$	.71	.69	.66	.64	.84	.80	.62	.42	.69	.73
SPRT: $\delta = .15$	.71	.60	.53	.70	.85	.72	.67	.56	.69	.79
SPRT: $\delta = .20$	.75	.61	.51	.76	.82	.72	.67	.60	.67	.75
SPRT: $\delta = .25$	.79	.64	.78	.79	.86	.93	.72	.22	.72	1.00
SPRT: $\delta = .30$	.67	.59	.54	.70	.78	.70	.60	.68	.79	.90
SCSPRT: $\delta = .20; \gamma = .95$	.80	.60	.59	.43	.61	.79	.78	.60	.79	.98
SCSPRT: $\delta = .20; \gamma = .86$	.61	.65	.56	.76	.78	.75	.57	.54	.74	.86
SCSPRT: $\delta = .20; \gamma = .68$	.67	.66	.62	.68	.79	.69	.53	.62	.76	.84
SCSPRT: $\delta = .20; \gamma = .38$	.66	.60	.50	.72	.78	.72	.58	.60	.73	.86
SCSPRT: $\delta = .20; \gamma = .00$	.68	.48	.44	.70	.92	.69	.72	.44	.69	.82
GLR: $\delta = .10$	.77	.64	.53	.68	.76	.72	.60	.63	.67	.88
GLR: $\delta = .15$	.74	.66	.57	.68	.72	.78	.63	.52	.73	.88
GLR: $\delta = .20$	.70	.76	.55	.70	.78	.83	.70	.40	.63	.92
GLR: $\delta = .25$	.59	.56	.52	.57	.88	.62	.60	.61	.80	.96
GLR: $\delta = .30$	.74	.64	.56	.66	.78	.73	.64	.55	.70	.88

Table A.9: Test length averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at the nearest cut-point, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\max} = .1$. The midpoint between each of the vertical bars is close to a classification bound, and only a few θ_i on either side of each classification bound are presented for clarity.

Condition	$\theta_i = -1.6$	-1.5	-1.4	-1.3	-1.2	-0.7	-0.6	-0.5	-0.4	-0.3
SPRT: $\delta = .10$	20.9	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .15$	20.9	20.9	20.9	21.0	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .20$	19.5	19.5	20.2	20.7	21.0	21.0	21.0	21.0	21.0	21.0
SPRT: $\delta = .25$	18.7	18.3	20.6	19.6	19.5	20.6	20.8	20.8	20.8	20.2
SPRT: $\delta = .30$	16.1	18.6	16.9	18.0	19.8	19.1	19.2	20.2	19.6	19.5
SCSPRT: $\delta = .20; \gamma = .95$	16.2	17.6	18.0	18.8	18.9	19.1	19.3	19.4	19.2	19.4
SCSPRT: $\delta = .20; \gamma = .86$	16.7	18.0	17.4	18.0	18.0	18.7	19.0	19.1	19.2	19.2
SCSPRT: $\delta = .20; \gamma = .68$	17.3	17.7	18.0	18.1	18.1	18.5	18.8	18.9	19.1	19.1
SCSPRT: $\delta = .20; \gamma = .38$	16.8	17.0	17.8	18.0	17.6	17.9	18.4	18.5	18.4	18.6
SCSPRT: $\delta = .20; \gamma = .00$	14.1	15.2	15.7	16.7	16.5	17.8	17.6	17.9	17.4	17.2
GLR: $\delta = .10$	18.0	19.1	19.5	20.2	20.7	21.0	21.0	21.0	21.0	21.0
GLR: $\delta = .15$	18.0	20.1	20.0	20.7	20.4	21.0	20.9	20.9	21.0	21.0
GLR: $\delta = .20$	17.8	18.1	19.6	19.4	20.0	20.6	20.8	20.9	20.9	20.9
GLR: $\delta = .25$	17.0	16.7	19.1	19.5	19.6	20.4	20.6	20.3	20.5	20.5
GLR: $\delta = .30$	16.7	17.6	17.4	18.4	18.5	19.6	19.6	19.6	19.7	19.7

Condition	$\theta_i = 0.1$	0.2	0.3	0.4	0.5	1.0	1.1	1.2	1.3	1.4
SPRT: $\delta = .10$	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	21.0	20.9
SPRT: $\delta = .15$	21.0	21.0	21.0	21.0	21.0	21.0	20.9	20.9	20.6	19.7
SPRT: $\delta = .20$	20.9	20.7	20.8	20.6	20.6	20.0	20.1	19.5	18.5	17.6
SPRT: $\delta = .25$	20.7	20.0	20.6	18.5	18.6	18.9	18.8	18.0	17.1	16.6
SPRT: $\delta = .30$	19.5	19.4	18.5	17.6	17.0	18.1	16.5	16.0	17.2	16.4
SCSPRT: $\delta = .20; \gamma = .95$	19.3	18.9	19.2	18.8	18.2	17.9	17.9	17.8	17.7	17.2
SCSPRT: $\delta = .20; \gamma = .86$	19.3	19.3	18.7	18.7	18.4	17.3	17.0	17.6	16.4	15.2
SCSPRT: $\delta = .20; \gamma = .68$	18.9	18.7	18.4	18.2	17.9	17.2	16.7	17.0	16.6	15.2
SCSPRT: $\delta = .20; \gamma = .38$	18.2	18.2	17.8	17.6	17.2	16.0	17.2	16.3	16.1	14.9
SCSPRT: $\delta = .20; \gamma = .00$	18.1	17.6	17.2	17.4	16.5	15.9	15.8	15.7	15.7	13.0
GLR: $\delta = .10$	21.0	21.0	20.8	20.9	20.6	20.3	19.9	19.7	18.6	16.8
GLR: $\delta = .15$	21.0	20.9	20.4	20.9	20.5	19.7	19.3	19.2	17.8	16.4
GLR: $\delta = .20$	20.8	20.8	20.5	20.1	19.3	18.6	19.1	19.0	17.1	15.1
GLR: $\delta = .25$	20.4	20.1	19.9	19.3	18.0	17.1	16.3	16.8	16.8	14.7
GLR: $\delta = .30$	19.0	18.5	19.1	17.7	17.4	15.6	15.0	15.8	15.3	13.7

Table A.10: Classification accuracy averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at the nearest cut-point, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\max} = .1$. The midpoint between each of the vertical bars is close to a classification bound, and only a few θ_i on either side of each classification bound are presented for clarity.

Condition	$\theta_i = -1.6$	-1.5	-1.4	-1.3	-1.2	-0.7	-0.6	-0.5	-0.4	-0.3
SPRT: $\delta = .10$	.75	.70	.49	.65	.71	.73	.68	.50	.60	.71
SPRT: $\delta = .15$	.66	.72	.70	.82	.84	.65	.69	.42	.53	.74
SPRT: $\delta = .20$	.73	.67	.53	.62	.83	.77	.61	.52	.64	.74
SPRT: $\delta = .25$	.78	.56	.42	.63	.79	.76	.78	.72	.72	.71
SPRT: $\delta = .30$	.84	.60	.60	.80	.94	.93	.70	.64	.55	.75
SCSPRT: $\delta = .20; \gamma = .95$	.72	.60	.52	.69	.74	.76	.56	.52	.62	.71
SCSPRT: $\delta = .20; \gamma = .86$	.70	.62	.54	.58	.76	.74	.60	.55	.61	.81
SCSPRT: $\delta = .20; \gamma = .68$	.72	.57	.39	.62	.79	.72	.66	.47	.60	.69
SCSPRT: $\delta = .20; \gamma = .38$	.70	.63	.54	.68	.73	.75	.64	.53	.52	.65
SCSPRT: $\delta = .20; \gamma = .00$	.70	.71	.52	.57	.78	.86	.58	.54	.56	.72
GLR: $\delta = .10$	.76	.65	.50	.67	.71	.78	.62	.46	.64	.74
GLR: $\delta = .15$	.83	.74	.55	.42	.78	.82	.81	.61	.68	.88
GLR: $\delta = .20$	.72	.67	.49	.71	.78	.80	.57	.52	.56	.80
GLR: $\delta = .25$	.71	.62	.39	.62	.73	.76	.72	.58	.57	.75
GLR: $\delta = .30$	.73	.59	.50	.62	.77	.86	.64	.55	.57	.69

Condition	$\theta_i = 0.1$	0.2	0.3	0.4	0.5	1.0	1.1	1.2	1.3	1.4
SPRT: $\delta = .10$	.72	.62	.50	.70	.80	.81	.68	.52	.74	.84
SPRT: $\delta = .15$	.74	.27	.61	.81	.78	.68	.55	.52	.65	.86
SPRT: $\delta = .20$	.76	.50	.47	.82	.86	.74	.73	.47	.75	.87
SPRT: $\delta = .25$	.80	.46	.47	.80	.77	.80	.61	.62	.79	.91
SPRT: $\delta = .30$	.70	.84	.66	.71	.94	.84	.41	.69	.65	.80
SCSPRT: $\delta = .20; \gamma = .95$	.83	.54	.51	.75	.83	.76	.67	.48	.73	.78
SCSPRT: $\delta = .20; \gamma = .86$	.72	.68	.62	.70	.80	.78	.60	.48	.74	.86
SCSPRT: $\delta = .20; \gamma = .68$	.68	.64	.58	.76	.85	.81	.62	.60	.76	.89
SCSPRT: $\delta = .20; \gamma = .38$	.74	.57	.50	.67	.79	.67	.64	.52	.74	.84
SCSPRT: $\delta = .20; \gamma = .00$	.60	.60	.58	.55	.89	.66	.64	.54	.76	.91
GLR: $\delta = .10$	.75	.58	.52	.63	.78	.76	.68	.51	.71	.82
GLR: $\delta = .15$	.75	.68	.55	.78	.78	.72	.56	.64	.75	.89
GLR: $\delta = .20$	.75	.63	.49	.65	.76	.83	.60	.53	.79	.87
GLR: $\delta = .25$	.77	.67	.49	.64	.86	.78	.60	.58	.67	.88
GLR: $\delta = .30$	.74	.64	.61	.68	.74	.73	.65	.63	.71	.90

Appendix B

Tables: Aggregate over Distribution

The following tables indicate the main effects and corresponding effect sizes from ANOVAs predicting mean test length and classification accuracy from several predictors including stopping rule, item selection, ability estimation, exposure control, and various two-way interactions.

Table B.1: The average percentage classified correctly and number of items administered aggregated within each termination criterion.

Condition		Avg. Test Length	Avg. Classification Accuracy
SPRT:	$\delta = .10$	20.9	.771
SPRT:	$\delta = .15$	20.5	.775
SPRT:	$\delta = .20$	19.7	.772
SPRT:	$\delta = .25$	18.5	.772
SPRT:	$\delta = .30$	17.1	.770
SCSPRT:	$\delta = .20; \gamma = .95$	17.8	.769
SCSPRT:	$\delta = .20; \gamma = .86$	17.5	.768
SCSPRT:	$\delta = .20; \gamma = .68$	17.2	.769
SCSPRT:	$\delta = .20; \gamma = .38$	16.7	.765
SCSPRT:	$\delta = .20; \gamma = .00$	15.8	.760
SCSPRT:	$\delta = .25; \gamma = .95$	17.3	.771
SCSPRT:	$\delta = .25; \gamma = .86$	17.2	.769
SCSPRT:	$\delta = .25; \gamma = .68$	16.8	.770
SCSPRT:	$\delta = .25; \gamma = .38$	16.4	.764
SCSPRT:	$\delta = .25; \gamma = .00$	15.7	.757
GLR:	$\delta = .10$	19.6	.769
GLR:	$\delta = .15$	19.2	.770
GLR:	$\delta = .20$	18.6	.771
GLR:	$\delta = .25$	17.7	.768
GLR:	$\delta = .30$	16.6	.763

Table B.2: The average percentage classified correctly and number of items administered aggregated within each item selection method.

SPRT/GLR Conditions

Condition		Avg. Test Length	Avg. Classification Accuracy
Maximum FI:	$\hat{\theta}_i$	18.9	.769
Maximum FI:	Bound	18.8	.771

SCSPRT Conditions

Condition		Avg. Test Length	Avg. Classification Accuracy
Maximum FI:	$\hat{\theta}_i$	16.8	.764
Maximum FI:	Bound	16.9	.769

Table B.3: The average percentage classified correctly and number of items administered aggregated within each ability estimation method.

SPRT/GLR Conditions

Condition	Avg. Test Length	Avg. Classification Accuracy
Maximum Likelihood Estimation	18.8	.769
Weighted Likelihood Estimation	18.8	.771

SCSPRT Conditions

Condition	Avg. Test Length	Avg. Classification Accuracy
Maximum Likelihood Estimation	16.8	.765
Weighted Likelihood Estimation	16.9	.768

Table B.4: The average percentage classified correctly and number of items administered aggregated within each exposure control method.

SPRT/GLR Conditions

Condition		Avg. Test Length	Avg. Classification Accuracy
No Exposure Control		18.6	.778
Sympson-Hetter:	$r_{\max} = .2$	18.7	.774
Sympson-Hetter:	$r_{\max} = .1$	19.2	.758

SCSPRT Conditions

Condition		Avg. Test Length	Avg. Classification Accuracy
No Exposure Control		16.6	.773
Sympson-Hetter:	$r_{\max} = .2$	16.7	.771
Sympson-Hetter:	$r_{\max} = .1$	17.2	.755

Table B.5: The sums of squares and $\eta^2 = \frac{SSF}{SST}$, where SSF is the sums of squares of a particular factor, for an ANOVA predicting mean classification accuracy for those termination conditions resulting in an average of 17 or more items per CAT. The ANOVA was run with all main effects and those interactions that relate to the termination factor.

Variance Type	Sums of Squares	η^2
Exposure	0.01322	.678
Termination	0.00054	.028
Term by Expos	0.00074	.038
Term by Estim	0.00024	.012
Item Selection	0.00033	.017
Term by Select	0.00037	.019
Ability Estimation	0.00012	.006
Residuals	0.00392	
Total	0.01949	

Table B.6: The sums of squares and $\omega^2 = \frac{SSF - df_F \times MSE}{SST + MSE}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting mean test length. The ANOVA was run with all main effects and those interactions that relate to the termination factor.

Variance Type	Sums of Squares	ω^2
Termination	506.50	.958
Exposure	16.80	.032
Term by Expos	3.73	.007
Term by Estim	0.67	.001
Term by Select	0.58	.001
Ability Estimation	0.27	.001
Item Selection	0.01	.000
Residuals	0.27	
Total	528.84	

Table B.7: The sums of squares and $\omega^2 = \frac{SSF - df_F \times MSE}{SST + MSE}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting mean classification accuracy. The ANOVA was run with all main effects and those interactions that relate to the termination factor.

Variance Type	Sums of Squares	ω^2
Exposure	0.01666	.565
Termination	0.00434	.125
Term by Expos	0.00121	.000
Term by Estim	0.00061	.000
Item Selection	0.00057	.018
Term by Select	0.00046	.000
Ability Estimation	0.00043	.014
Residuals	0.00505	
Total	0.02933	

Appendix C

Figures: Conditional on Ability

The following figures depict the conditional accuracy and test length for various conditions across a variety of ability values. In all of the figures, a Sympson-Hetter item exposure control method was implemented, as described in Section 3.4. Note that even though many of the stopping rules did not appear to accurately classify simulees, much of the messiness is due to replicating the classification task only 400 times.

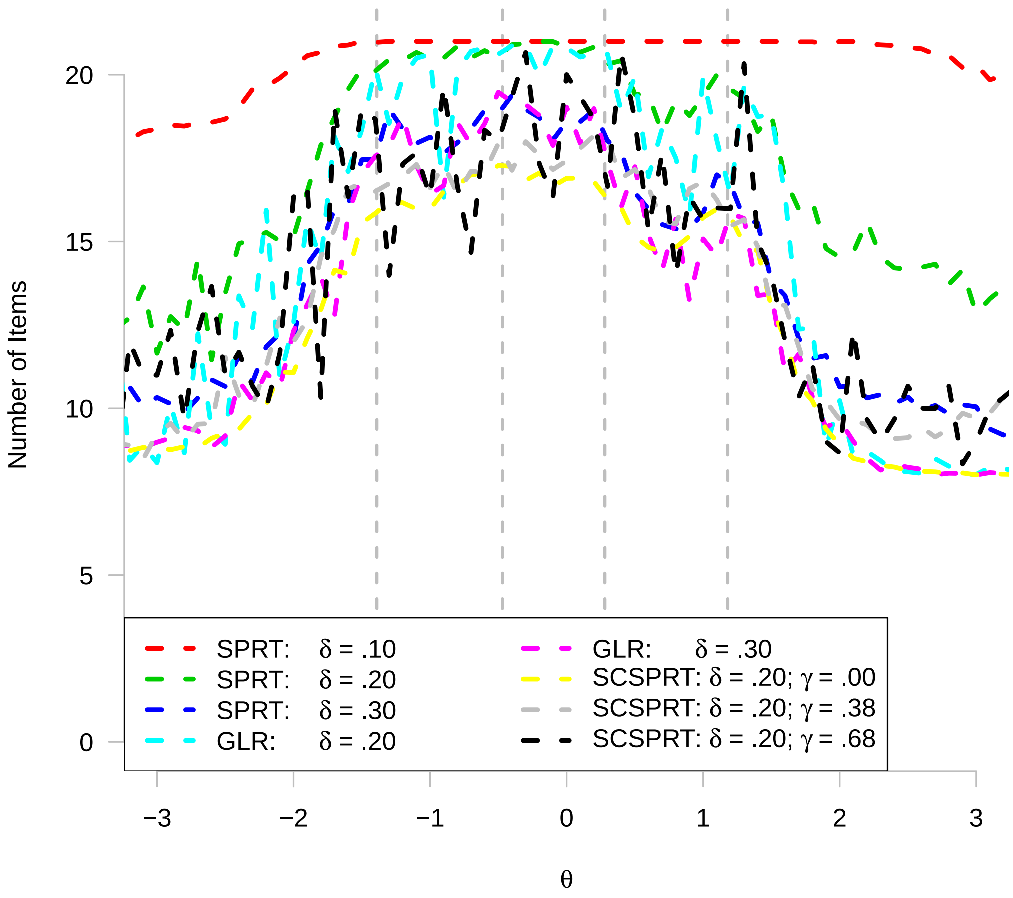


Figure C.1: Test length averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at $\hat{\theta}_i$, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\max} = .2$. The vertical bars represent the classification bounds. Only a few termination conditions are presented for illustration purposes.

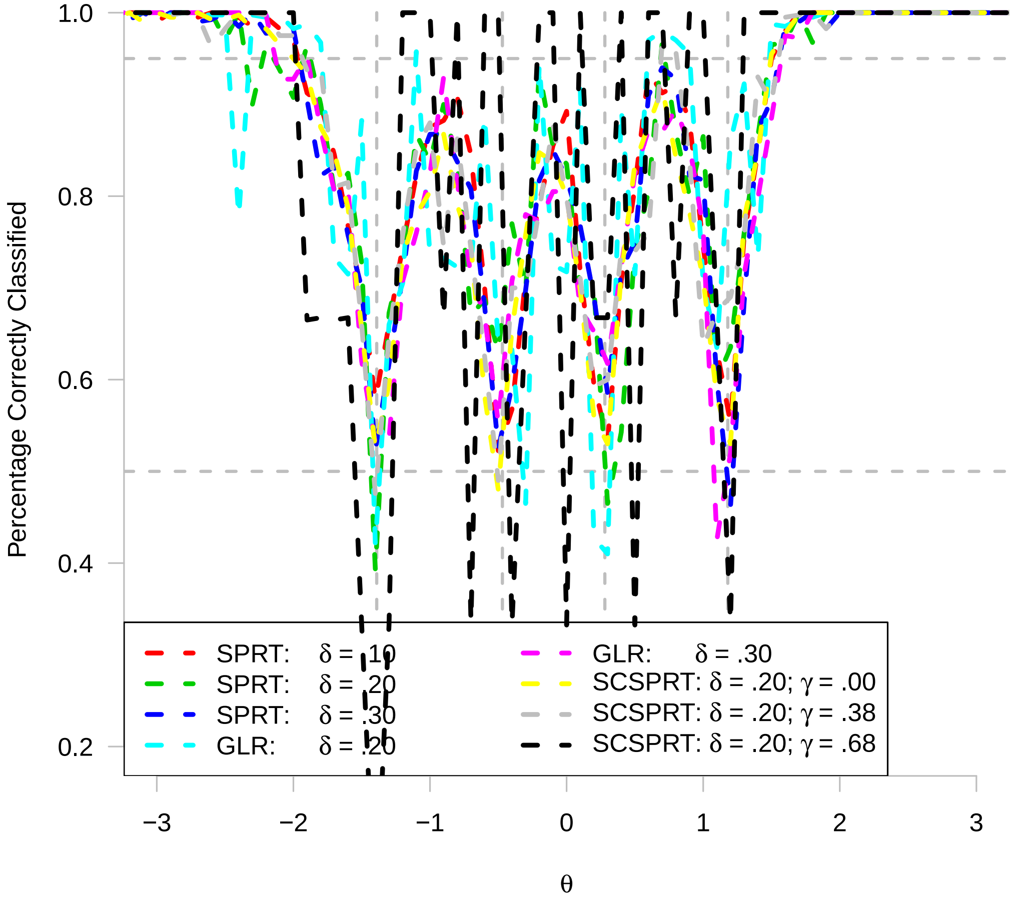


Figure C.2: Classification accuracy averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at $\hat{\theta}_i$, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\max} = .2$. The vertical bars represent the classification bounds, and the horizontal bars 50% classification accuracy and 95% classification accuracy. Only a few termination conditions are presented for illustration purposes.

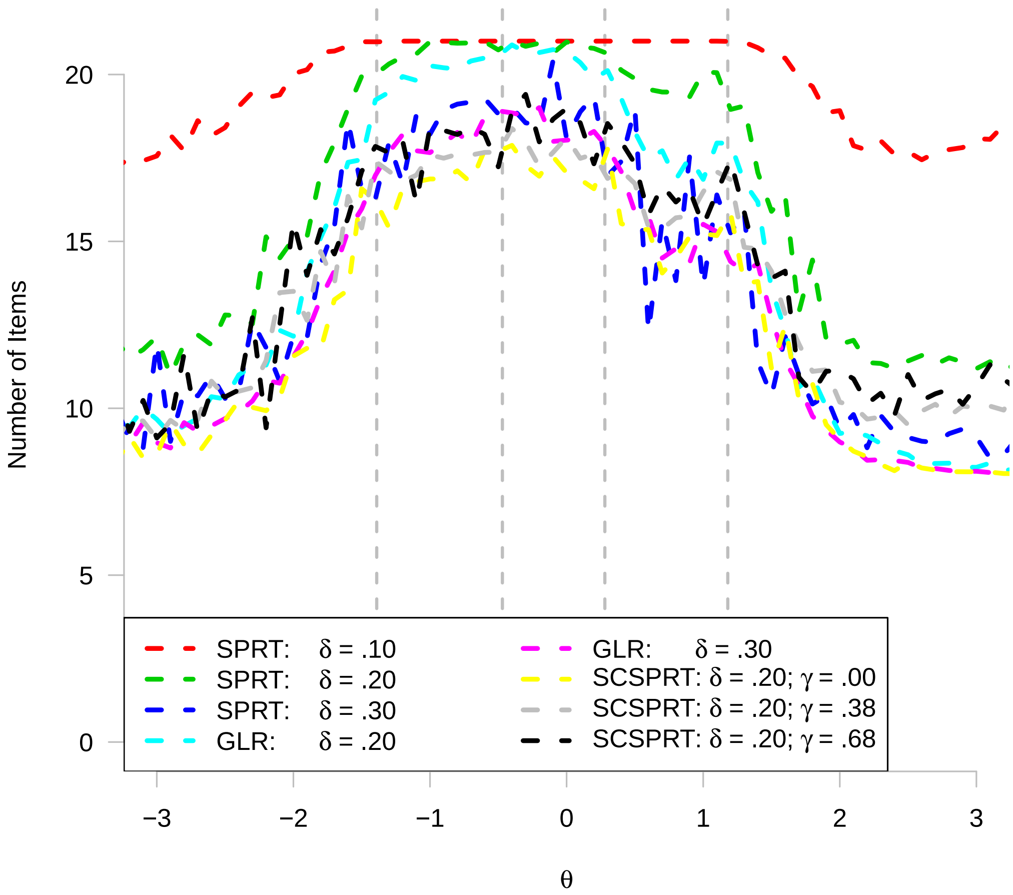


Figure C.3: Test length averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at the nearest cut-point, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\max} = .2$. The vertical bars represent the classification bounds. Only a few termination conditions are presented for illustration purposes.

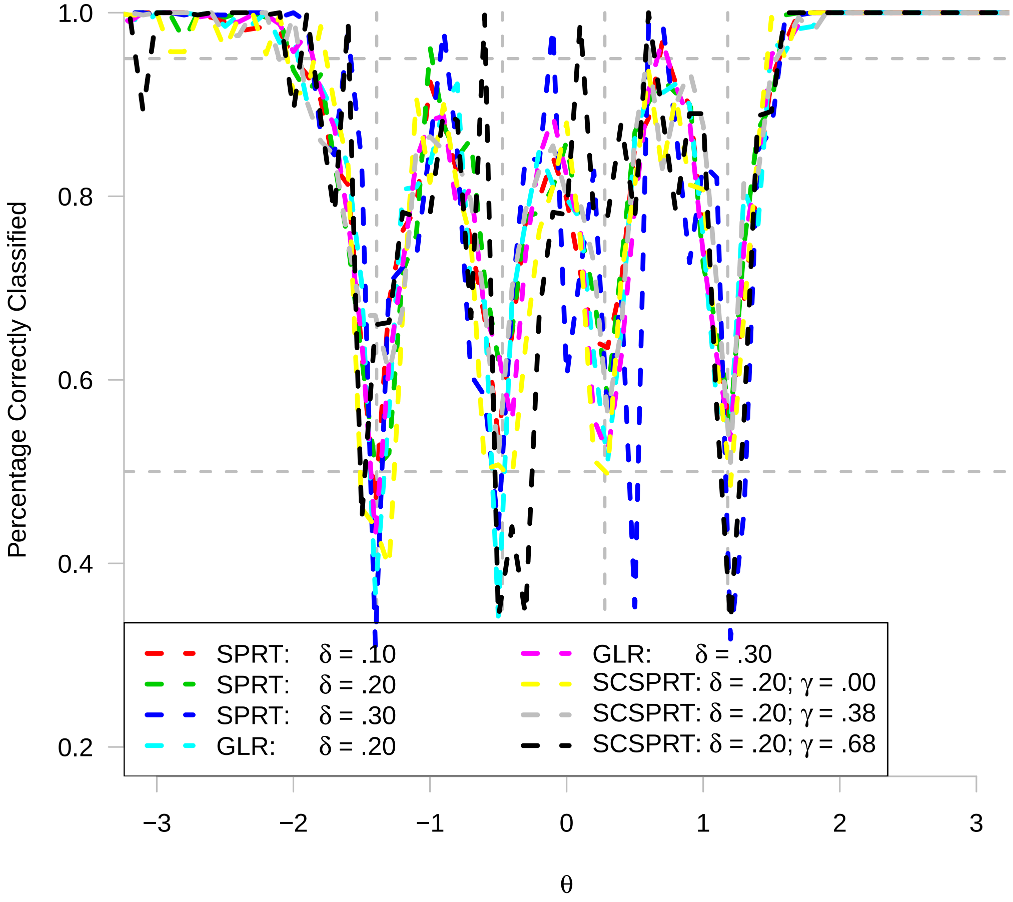


Figure C.4: Classification accuracy averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at the nearest cut-point, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\max} = .2$. The vertical bars 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.

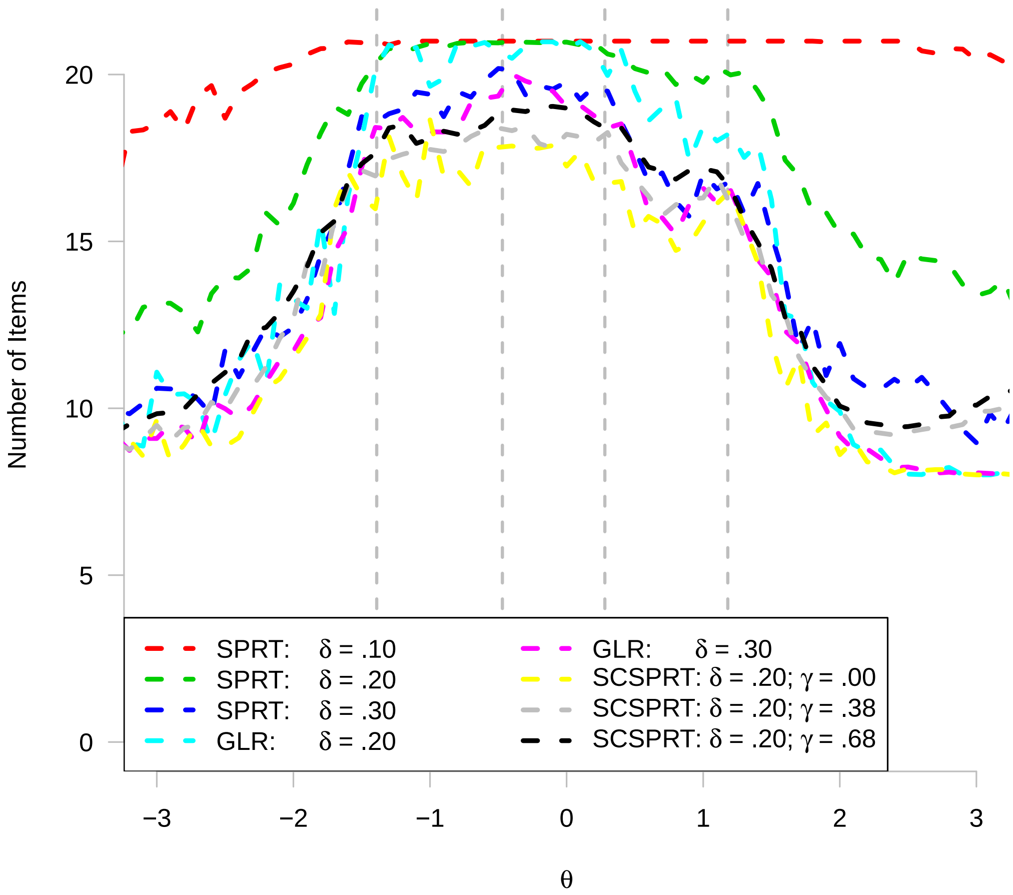


Figure C.5: Test length averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at $\hat{\theta}_i$, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\max} = .1$. The vertical bars represent the classification bounds. Only a few termination conditions are presented for illustration purposes.

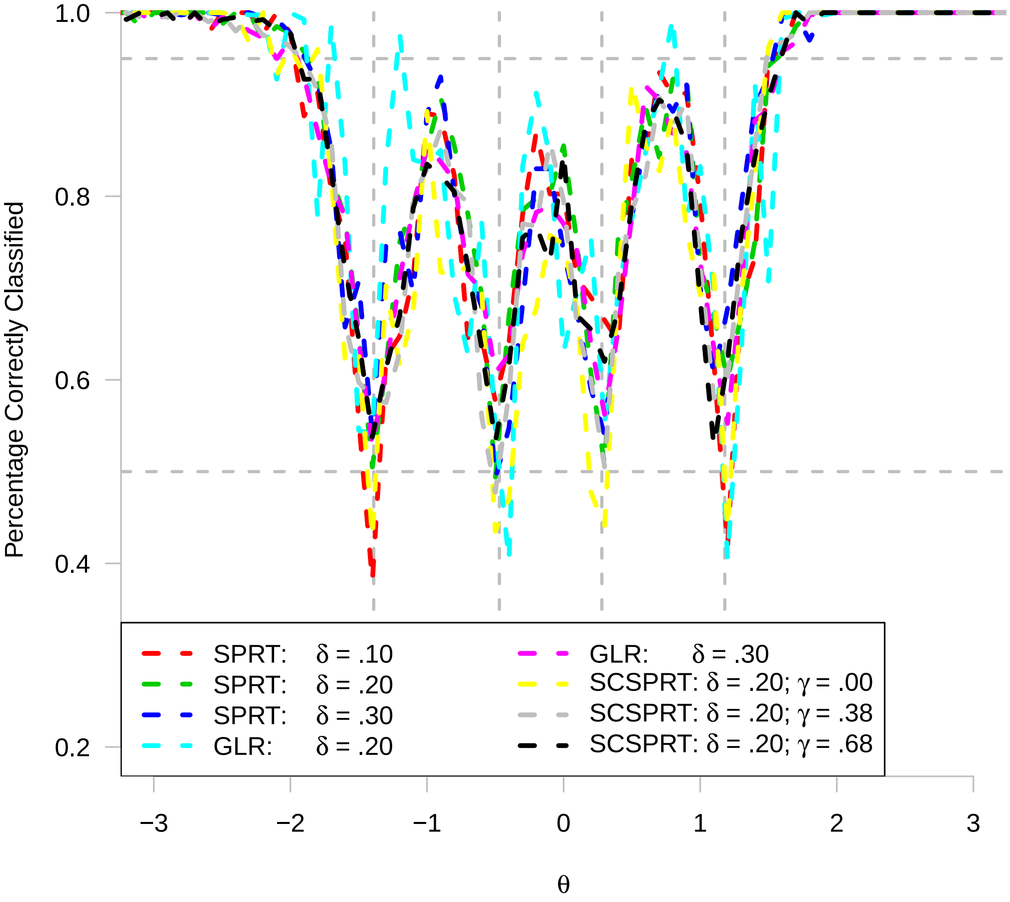


Figure C.6: Classification accuracy averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at $\hat{\theta}_i$, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\max} = .1$. The vertical bars represent the classification bounds, and the horizontal bars 50% classification accuracy and 95% classification accuracy. Only a few termination conditions are presented for illustration purposes.

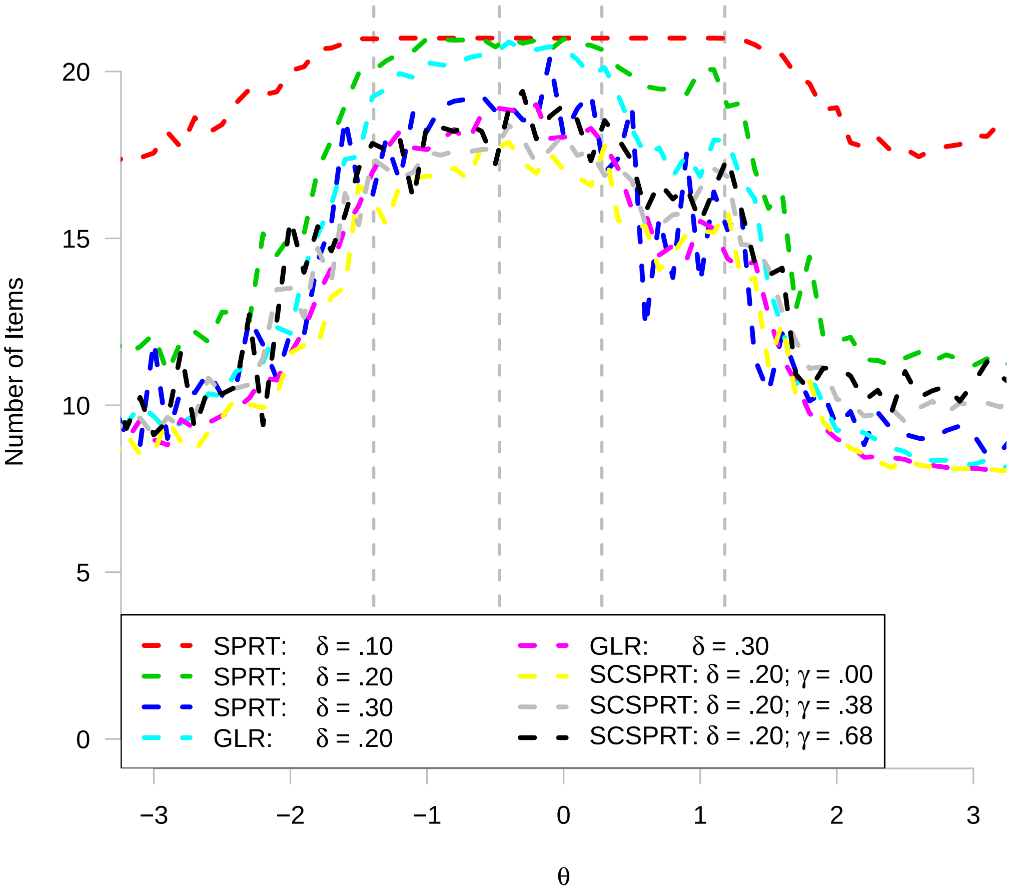


Figure C.7: Test length averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at the nearest cut-point, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\max} = .1$. The vertical bars represent the classification bounds. Only a few termination conditions are presented for illustration purposes.

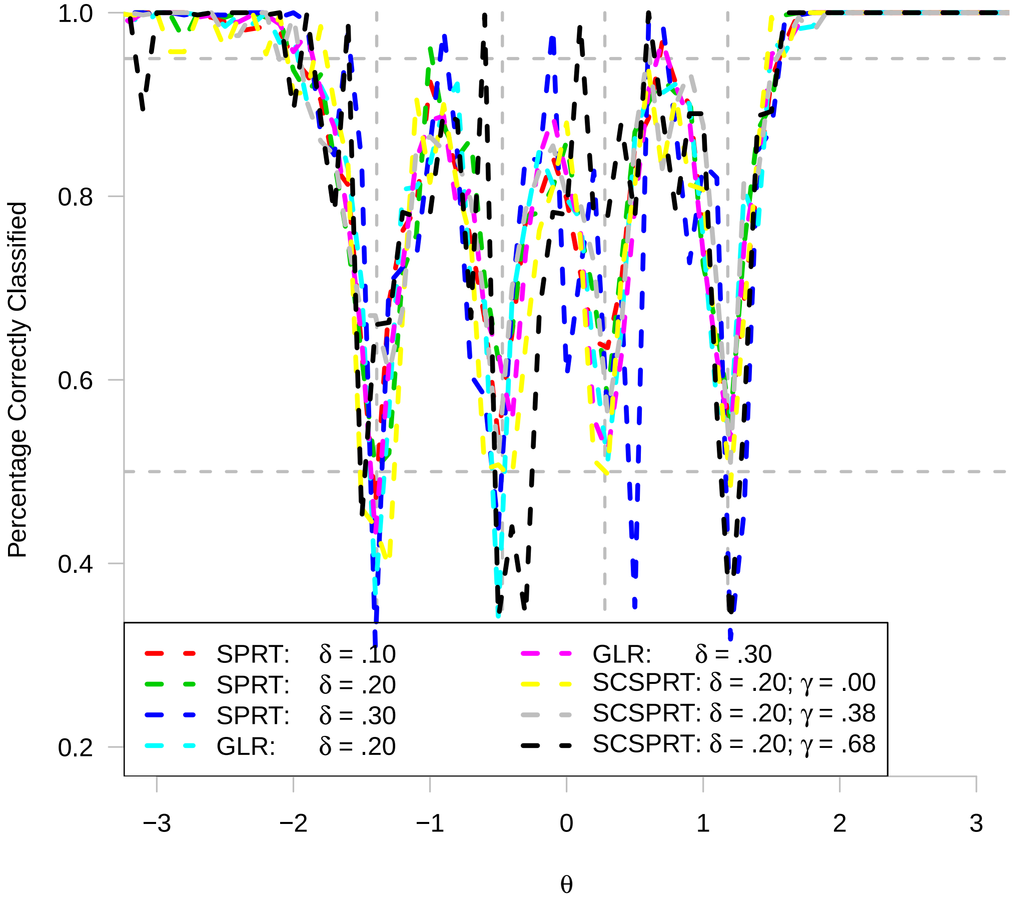


Figure C.8: Classification accuracy averaged over 400 classification CATs conditional on particular values of θ_i with items selected by Fisher information at the nearest cut-point, ability estimated by maximum likelihood estimation, and an item exposure control of $r_{\max} = .1$. The vertical bars 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.

Appendix D

Figures: Aggregate over Distribution

The following figures depict the relationship between accuracy and test length for various conditions aggregated across a distribution of simulees. In all of the figures, a Sympson-Hetter item exposure control method was implemented, as described in Section 3.4, and ability was simulated from a standard normal distribution.

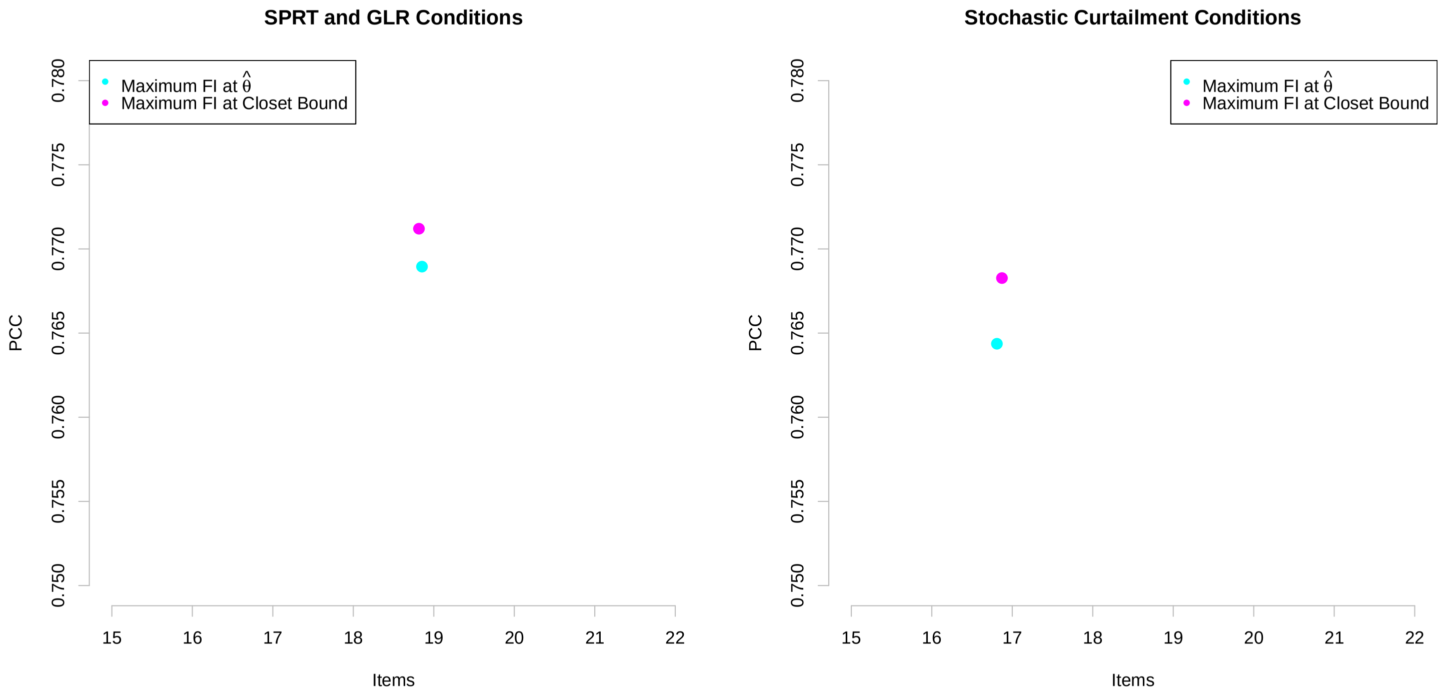


Figure D.1: Side-by-side scatterplots of the average percentage classified correctly by number of items administered based on each item selection method. The left plot contains all of the SPRT and GLR conditions, and the right plot contains all of the SCSPRT conditions. Points are color coded according to item selection method.

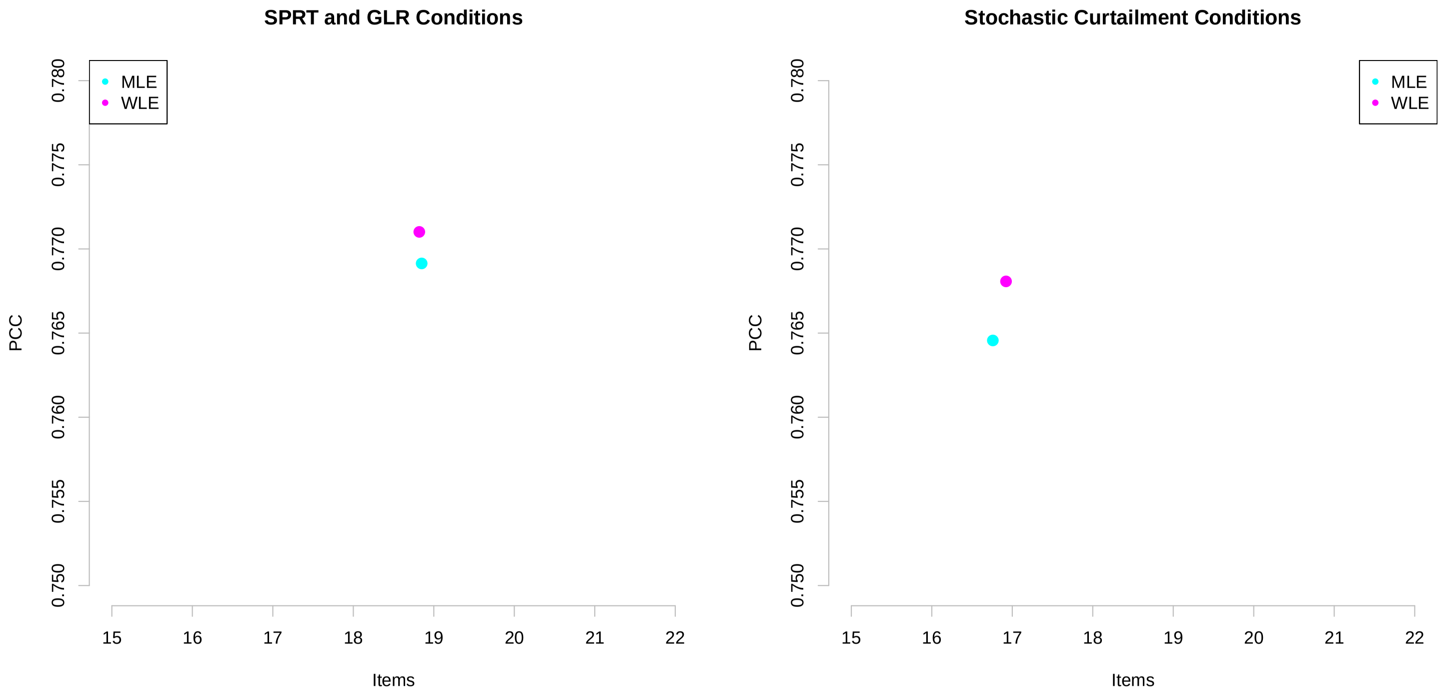


Figure D.2: Side-by-side scatterplots of the average percentage classified correctly by number of items administered based on each termination criterion. The left plot contains all of the SPRT and GLR conditions, and the right plot contains all of the SCSPRT conditions. Points are color coded according to ability estimation method.

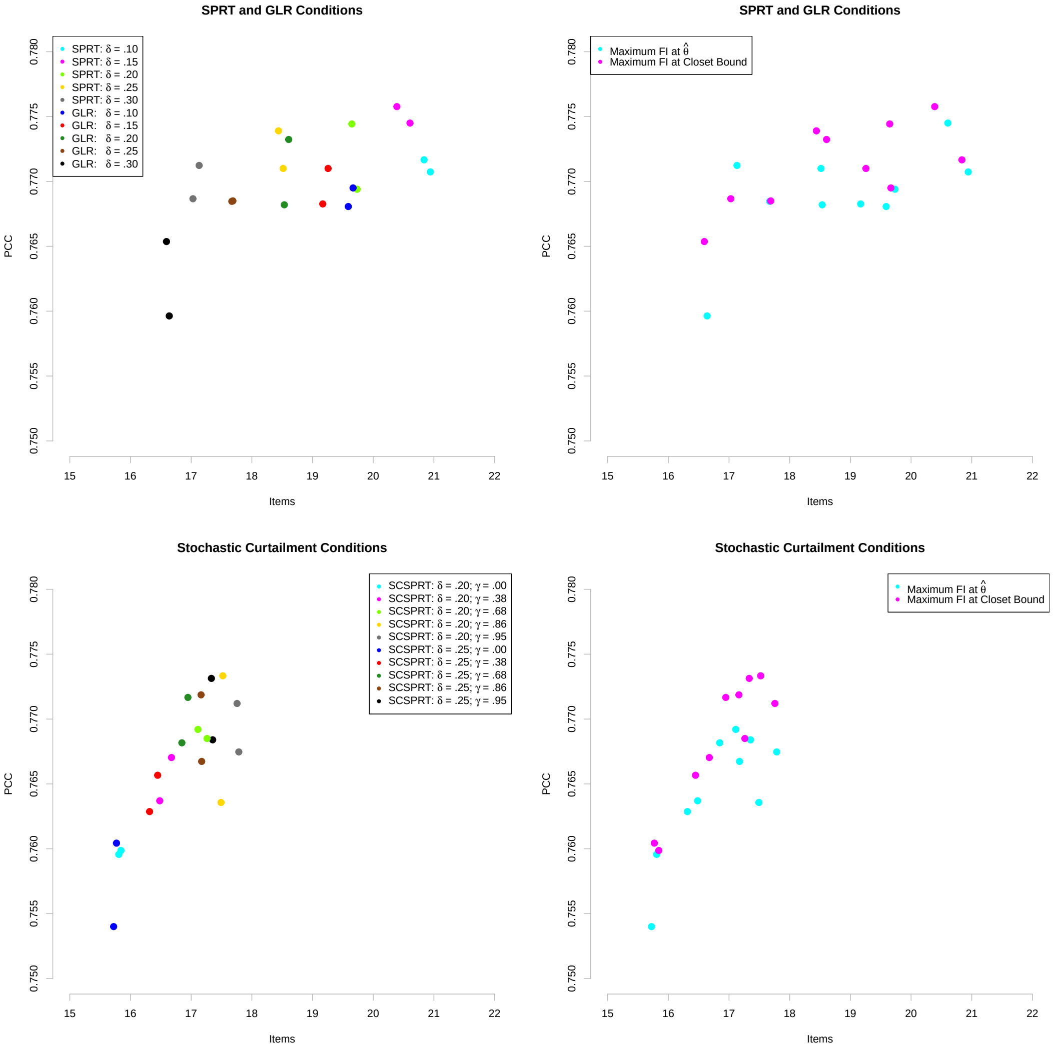


Figure D.3: Side-by-side scatterplots of the average percentage classified correctly by number of items administered based on the interaction between item selection and termination criterion. The top plots contain all of the SPRT and GLR conditions, and the bottom plots contain all of the SCSPRT conditions. The left plots are color coded according to termination criterion, and the right plots are color coded according to item selection condition

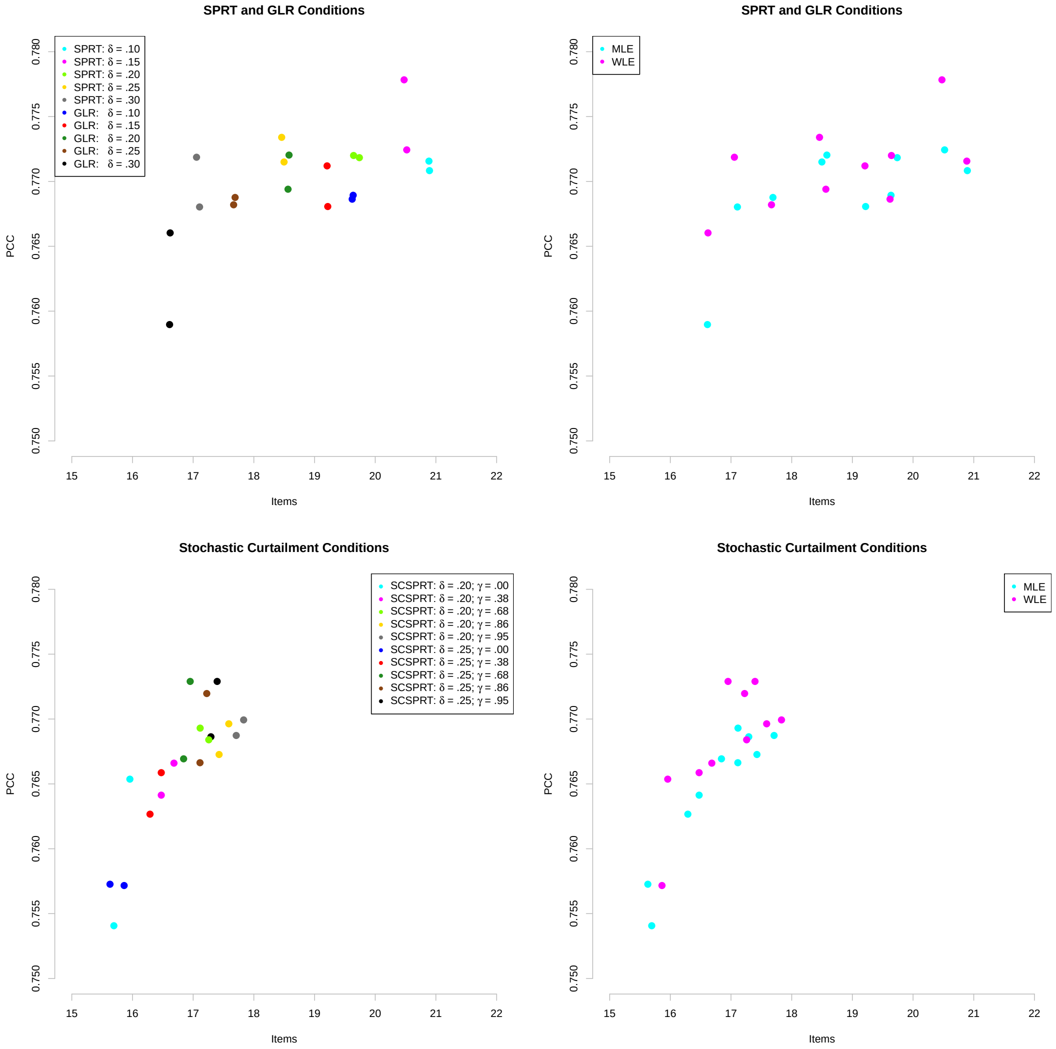


Figure D.4: Side-by-side scatterplots of the average percentage classified correctly by number of items administered based on the interaction between ability estimation and termination criterion. The top plots contain all of the SPRT and GLR conditions, and the bottom plots contain all of the SCSPRT conditions. The left plots are color coded according to termination criterion, and the right plots are color coded according to ability estimation method