A Comparison of Item Selection Methods and Stopping Rules in Multi-category Computerized Classification Testing

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

King Yiu Suen

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Dr. David J. Weiss, Advisor

December 2022

Abstract

Computerized classification testing (CCT) aims to classify people into one of two or more possible categories while maximizing accuracy and minimizing test length. Two key components of CCT are the item selection method and the stopping rule. The current study used simulation to compare the performance of various item selection methods and stopping rules for multi-category CCT in terms of average test length (ATL) and percentage of correct classifications (PCC) under a wide variety of conditions. Item selection methods examined include selecting items to maximize the Fisher information at the ability estimate, Fisher information at the nearest cutoff, and the sum of Fisher information of all cutoffs weighted with the likelihood function. The stopping rules considered were a multi-hypothesis sequential probability ratio test (mSPRT) and a multi-category generalized likelihood ratio test (mGLR), combined with three variations of stochastic curtailment methods (SC-Standard, SC-MLE and SC-CI). Manipulated conditions included the number of cutoffs, the distribution of the examinees' abilities, the width of the indifference region, the shape of the item bank information function, and whether the items were calibrated with estimation error. Results suggested that the combination of mGLR and SC-MLE consistently had the best balance of ATL and PCC. The three item selection methods performed similarly across all conditions.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Background

The purpose of classification tests is to classify individuals into mutually exclusive categories such as mastery or non-mastery. In a conventional paper-and-pencil test, all examinees receive the same number of items in a fixed order. With the advent of technology, tests can easily be delivered by computers. In a computerized classification test (CCT), items can be presented to an examinee one at a time and each item can be scored immediately after the examinee provides an answer. Sophisticated methods have been developed to decide whether the item responses collected at any point in the test provide adequate confidence to make a classification decision. Once an unambiguous decision is reached, the test can be terminated. As a result, examinees who are easier to be classified, such as those whose ability is far below or above the cutoff score, can receive fewer items. Variable-length testing can reduce test administration time and alleviate the respondent's burden (Bass et al., 2015; Finkelman et al., 2011; Gibbons et al., 2008). It also indirectly limits the exposure rate of the items, and hence upholds test security, which is especially important for high-stake tests (Huebner & Fina, 2015).

Classical test theory (e.g., Gulliksen, 1950) and item response theory (IRT; e.g., Embretson & Reise, 2013) are the two possible psychometric models that can be used as a basis for CCT. Most research in CCT has focused on IRT, because it offers a rich family of parametric models that describe the interaction between the characteristics of an item and the ability of an examinee (e.g., Birnbaum, 1968; Masters, 1982; Muraki, 1992; Rasch, 1960; Reckase, 2009; Samejima, 1969), an individualized standard error of measurement (Weiss, 2011), as well as a better solution to test construction (Lord, 1980),

the identification of biased items (Lord, 1980), adaptive testing (Weiss, 1983), and the equating of test scores (Cook & Eignor, 1983, 1989). Item banks for large-scale testing programs that can afford to apply CCT are often calibrated with IRT (Thompson, 2009a). For these reasons, this study focused on IRT and did not apply methods based on classical test theory (e.g., Frick, 1992; Rudner, 2002; Vos, 2000).

Several stopping rules have been proposed for two-category CCT. Ability confidence interval (ACI; Thompson, 2011; Weiss & Kingsbury, 1984) constructs a confidence interval (CI) around the ability (θ) estimate after each item. The test stops when the CI falls completely below or above the cutoff score. The sequential probability ratio test (SPRT; Wald, 1947) requires the specification of two points, one below and one above the cutoff score. The interval formed by these two points is known as the indifference region. Whether the test should be terminated depends on the ratio of the likelihoods of item responses at these two points. The generalized likelihood ratio test (GLR; Bartroff et al. 2008; Thompson, 2009b) makes the decision based on the ratio between the maximum likelihood for the parameter space below the lower endpoint of the indifference region and the maximum likelihood for the parameter space above the upper endpoint of the indifference region. Stochastic curtailment (SC; Finkelman, 2008) terminates the test once the probability that the tentative classification decision based on items administered at any point in the test would remain unchanged, had the test continued. The calculation of this probability involves computing the expected responses of future items. Finkelman (2008) originally proposed to evaluate this expectation at the endpoints of the indifference region (SC-Standard). Finkelman (2010) later found that the test efficiency can be further improved if the expectation was evaluated at the current θ

estimate (SC-MLE), at the endpoints of the CI around the θ estimate (SC-CI), or over the posterior distribution of the θ estimate (SC-Bayes). All of these stopping rules have been generalized to, and examined in, multi-category scenarios (Wang et al., 2021), except SC-MLE, SC-CI, and SC-Bayes. Multiple studies have called for more research to fill in this gap (e.g., Finkelman, 2010; Sie et al., 2012; Wang et al., 2021).

Earlier research has selected items randomly (e.g., Ferguson, 1969, Vos, 1998). This, however, does not make use of any information regarding the items or the examinees (Thompson, 2007b). A more intelligent approach would evaluate the unadministered items in the bank and decide which can best facilitate the classification decision. There are mainly two types of intelligent item selection methods, cutscore-based and estimate-based (Thompson, 2007a, 2007b). In a two-category CCT, cutscore-based methods select items that maximize the amount of information at the cutoff score (Spray & Reckase, 1994), or items that can best differentiate the two groups divided by the cutoff score (Eggen, 1999). Estimate-based methods aim to select items that maximize the amount of information at the examinee's θ estimate (Reckase, 1983). Depending on the stopping rule used, the appropriate type of item selection method can vary, as different stopping rules base their decision on different types of information. Thompson (2009a) suggested that SPRT should be paired with cutscore-based methods and ACI should be paired with estimate-based methods. However, there are stopping rules that requires the evaluation of the likelihood at both the θ estimate and cutoff score(s), such as GLR. While Thompson (2009a, 2011) believed that GLR would work better with cutscore-based methods, Wang et al. (2021) suspected the opposite. No

studies to date have compared the performance of GLR with different item selection methods.

Moreover, all prior studies designed their item banks by generating item parameters with values drawn directly from a specified distribution. However, in an applied setting, item banks are created by administering a set of items to a calibration sample and then using the responses to estimate the item parameters. This estimation process inherently introduces error into the item bank. Simulation studies have shown that item calibration errors could result in spuriously high values of test information, biased θ estimates and underestimated standard errors (Hambleton et al., 1993; Hambleton & Jones, 1994; Patton et al., 2013; van der Linden & Glas, 2000). However, most of these studies examined the impact of item calibration errors only on tests that aim to obtain a point estimate of θ. To date, only one study has examined the impact in two-category CCT (Patton et al., 2013). The CCT literature has largely ignored the presence of item calibration errors when creating and using item banks. As noted above, some item selection methods and stopping rules involve estimating θ and the associated standard error. Thus, it is plausible that item calibration errors could influence their performance.

As discussed above, previous research on CCT has mainly focused on classification into two categories. However, many assessments classify individuals into three or more categories. For example, occupational aptitude testing might require a classification of job applicants into inferior, mediocre and superior proficiencies (Gnambs & Batinic, 2011). The Beck Depression Inventory (Beck et al., 1961) classifies individuals into five levels of depression. Many statewide tests under No Child Left Behind also classify students into multiple proficiency groups (Finkelman, 2010).

Purposes

The general purpose of this study is to compare various combinations of item selection methods and stopping rules for multi-category CCT. Specifically, this study aimed to (1) examine whether SC-MLE and SC-CI are still more efficient than SC-Standard in a multi-category CCT, (2) explore which item selection method works best with GLR in multi-category CCT, and (3) examine the effects of item calibration errors in multi-category CCT. This study will provide test practitioners comprehensive and guiding information regarding several major components of a multi-category CCT, namely, stopping rules, item selection methods, and the item bank.

Chapter 2. Item Response Theory

In CCT, an appropriate mathematical model is needed to characterize the interaction between the examinee's θ and responses on test items. Item response theory (IRT) is commonly used for this purpose.

Assumptions

A common assumption of IRT models is that only one ability is measured by the items in a test (Hambleton et al., 1991). For example, in a math test, math ability should be the only factor that influences test performance. If a math test is administered to students in their secondary language, then this assumption might be violated, as the examinee's language skills will be also required to solve the problems. IRT models that satisfy this assumption are referred to as unidimensional models. Models that assume that more than one ability is necessary to account for examinee test performance are referred to as multidimensional. While multidimensional IRT models exist (Reckase, 2009), their application is beyond the scope of this study.

The second assumption is local independence. It means that conditional on an examinee's θ, the responses to any pair of items should be statistically independent (Hambleton et al., 1991). In other words, how an examinee responds to an item should solely depend on their θ, not by how they respond to any other item in the test. This assumption is usually violated if there are several items pertaining to the same reading passage or math story item, or if answers to later items depend on answers to earlier items. Research has shown that if the local independence assumption is violated, the precision of measurement might be overestimated (Sireci et al., 1991; Wainer & Thissen,

1996). Testlet response theory (Wainer et al., 2007) can be used as an alternative to IRT when there is a dependency among responses to a set of items.

Dichotomous Models

This study focused on CCT with dichotomous items. An item is said to be dichotomous when it has only two possible outcomes, for example, correct and incorrect. The three most widely used unidimensional dichotomous IRT models are the one-, two-, and three-parameter logistic models. As their names suggest, they differ in the number of parameters they incorporate.

Denote u_j as the response of item j by an examinee with ability θ, where $u_j = 1$ if it is answered correctly, and $u_j = 0$ otherwise. In the one-parameter logistic (1PL; Rasch, 1960) model, the probability of answering item j correctly by examinee i is defined as

$$P_j(\theta) = P\left(u_j = 1 \middle| \theta, b_j\right) = \frac{\exp\left(\theta - b_j\right)}{1 + \exp\left(\theta - b_j\right)} \quad (1)$$

The b_j parameter represents the difficulty of item j. If $\theta = b_j$, the probability of a correct response is exactly 0.5. The more difficult item j is, the higher the value of b_j, and hence the higher θ required to have a 50% chance of getting the item correct. It should be highlighted that Equation 1 reveals a very important feature of IRT: the examinee's ability θ and the difficulty parameter b_j lie on the same scale. A drawback of the 1PL model is that it has a constant discrimination parameter for all items. Also, in the 1PL model, an examinee with a very low θ has a close to zero probability of answering the item correctly. In practice, low-ability examinees might get the correct answer by guessing if they are given multiple-choice items (Hambleton et al., 1991).

The two-parameter logistic (2PL; Birnbaum, 1968) model is a generalization of the 1PL model that allows for differently discriminating items. The mathematical expression for the 2PL model is

$$P_j(\theta) = P\left(u_j = 1 \middle| \theta, a_j, b_j\right) = \frac{\exp\left[\mathrm{D}a_j\left(\theta - b_j\right)\right]}{1 + \exp\left[\mathrm{D}a_j\left(\theta - b_j\right)\right]}, \tag{2}$$

where a_j is the discrimination parameter for item j and $\mathrm{D} = 1.702$ is a scaling constant. A large value of a_j results in a larger difference in $P_j(\theta)$ between two examinees with θs in the vicinity of b_j. Items with a negative value of a_j are typically discarded, because it implies that $P_j(\theta)$ decreases as θ increases, which is clearly illogical (Hambleton et al., 1991). The reason for the existence of the scaling constant D is that the two-parameter IRT model originally developed by Lord (1952) was based on the cumulative normal distribution (normal ogive),

$$P_j(\theta) = \int_{-\infty}^{a_j(\theta - b_j)} \frac{1}{\sqrt{2\pi}} \exp(-t/2)\, dt \tag{3}$$

Birnbaum (1968) later discovered that with the use of the scaling constant, the logistic function in Equation 2 can approximate the normal ogive function in Equation 3 very closely (Camilli, 1994). Since the logistic form does not involve integration, it is more convenient to work with (Hambleton et al., 1991) and more commonly used. If $a_j = 1$ for all items, the 2PL is reduced to the 1PL model.

The three-parameter logistic (3PL; Birnbaum, 1968) model is a generalization of the 2PL model that allows for guessing behavior. The mathematical expression for the 3PL model is

$$P_j(\theta) = P\left(u_j = 1 \middle| \theta, a_j, b_j, c_j\right) = c_j + \left(1 - c_j\right) \frac{\exp\left[\mathrm{D}a_j\left(\theta - b_j\right)\right]}{1 + \exp\left[\mathrm{D}a_j\left(\theta_i - b_j\right)\right]}. \tag{4}$$

The additional parameter c_j is called the guessing parameter. It represents the probability of low-ability examinees answering the item correctly. When $c_j = 0$ for all items, the 3PL model reduces to the 2PL model.

Note that since item j is dichotomous, the probability of an incorrect response is simply

$$Q_j(\theta) = 1 - P_j(\theta). \tag{5}$$

Also note that Equations 1, 2 and 4 are sometimes known as item response functions (IRF).

θ Estimation Methods

The score of an IRT-based test is an estimate of θ. The estimation of θ is based on the likelihood function of the responses. Let $\boldsymbol{u} = (u_1, \ldots, u_J)$ be a vector of J item responses by an examinee. The likelihood function of a response pattern is defined as the joint probability of observing the responses. If the local independence assumption is met, the joint probability is simply the product of the probabilities associated with the responses to each individual item,

$$L(\theta|\boldsymbol{u}) = P(\boldsymbol{u}|\theta) = P\left(u_1, \ldots, u_J \middle| \theta\right) = \prod_{j=1}^{J} P_j(\theta)^{u_j} Q_j(\theta)^{1-u_j}, \tag{6}$$

A common θ estimation method is called maximum likelihood estimation (MLE). In MLE, the θ estimate ($\hat{\theta}$) is the θ value that maximizes the likelihood function,

$$\hat{\theta}^{\text{MLE}} = \arg\max_{\theta} L(\theta|\boldsymbol{u}). \tag{7}$$

As the number of items increases, the product of the probabilities in the likelihood function will potentially become very close to 0, and it will become difficult to represent

in a computer (De Ayala, 2013). Therefore, it is often recommended to work with the logarithm of the likelihood function instead (De Ayala, 2013),

$$\hat{\theta}^{\text{MLE}} = \arg\max_{\theta} \log L(\theta|\boldsymbol{u}) = \sum_{j=1}^{J}[u_j \log P_j(\theta) + (1-u_j)\log Q_j(\theta)], \quad (8)$$

Since the logarithm is a monotonically increasing function, the $\hat{\theta}^{\text{MLE}}$ found in Equation 7 will be the same as the $\hat{\theta}^{\text{MLE}}$ found in Equation 8. There is no closed-form solution to Equation 8. It must be solved by a numerical method such as the Newton's method (De Ayala, 2013). $\hat{\theta}^{\text{MLE}}$ is called the maximum likelihood estimate (also abbreviated as MLE).

One disadvantage of MLE is that for response patterns with all items correct or all incorrect, no finite MLE exists. This problem tends to occur at the beginning of a CCT, when only a few items have been administered. It can be overcome by implementing a Bayesian estimation method temporarily until a mixed response pattern is observed. The basic idea of Bayesian methods is to modify the likelihood function to incorporate any prior information we may have about θ. In particular, Bayesian methods require the specification of a prior distribution $\pi(\theta)$, the assumed distribution of θ before any response is observed. This can be based on previous experience or by making assumptions. For example, we can assume that θ is normally distributed with a specific mean and variance. Using the Bayes' rule, the conditional probability of θ given the responses $\boldsymbol{u}$, can be computed by

$$\pi(\theta|\boldsymbol{u}) = \frac{\pi(\theta)L(\theta|\boldsymbol{u})}{P(\boldsymbol{u})} = \frac{\pi(\theta)L(\theta|\boldsymbol{u})}{\int_{\theta} P(\theta)\, L(\theta|\boldsymbol{u})d\theta} \quad (9)$$

Equation 9 is known as the posterior distribution, the distribution of θ after taking the observed responses into account. The posterior distribution can be used to estimate θ.

The maximum a posteriori (MAP) method uses the mode of the posterior distribution as the $\hat{\theta}$,

$$\hat{\theta}^{\text{MAP}} = \arg\max_{\theta} \pi(\theta|\boldsymbol{u}), \tag{10}$$

whereas the expected a posteriori (EAP) method uses the expected value of the posterior distribution as the $\hat{\theta}$,

$$\hat{\theta}^{\text{EAP}} = \int_{\theta} \theta\pi(\theta|\boldsymbol{u})d\theta. \tag{11}$$

The integration can be approximated by using Gauss-Hermite quadrature (Stroud & Sechrest, 1966). Also, $\hat{\theta}$ from MAP or EAP can be biased, if the prior mean does not match the examinee's true θ (De Ayala, 2013; Wang, 2015).

Fisher Information and Standard Error of Measurement

In addition to obtaining a point estimate of θ, it is equally important to quantify how certain we are about an examinee's θ. In IRT, the extent to which an item can differentiate two nearby θ levels can be measured by Fisher information (FI; Weiss, 2011). Mathematically, the FI of item j is defined as the squared slope of the IRF divided by its variance, both conditional on θ,

$$I_j(\theta) == \frac{\left[\frac{\partial}{\partial\theta}P_j(\theta)\right]^2}{P_j(\theta)Q_j(\theta)}. \tag{12}$$

In the 3PL model, information is higher when the discrimination parameter a_j is higher, the difficulty parameter b_j is closer to θ, and the guessing parameter c_j is lower (Hambleton et al., 1991). Note that information is a function of θ, which means an item might provide considerable information at one θ level, but very little information at another. Birnbaum (1968) shows that an item provides its maximum information at

$$\theta = b_j + \frac{1}{Da_j}\log\left[0.5\left(1+\sqrt{1+8c_j}\right)\right]. \quad (13)$$

It can be observed from Equation 13 that if $c_j = 0$, the item provides its maximum information at $\theta = b_j$; if $c_j > 0$, the item provides its maximum information at a θ level higher than b_j. FI is additive, in that the FI of a test of J items is the sum of FI of those J items,

$$I(\theta) = \sum_{j=1}^{J} I_j(\theta)\,. \quad (14)$$

This indicates that items contribute independently to the test information function (TIF; Hambleton et al., 1991).

A concept related to FI is the standard error of measurement (SEM), which quantifies the amount of measurement error. The theoretical SEM can be obtained by taking the reciprocal of the square root of FI,

$$\text{SEM}(\theta) = \frac{1}{\sqrt{I(\theta)}} \quad (15)$$

Note that SEM, just as information, is a function of θ.

The FI described so far is called theoretical FI. It is in contrast to the observed FI, which is a variant of FI that is based on observed responses. The observed FI for an item is defined as the negative of the second derivative of the log-likelihood,

$$I_j(\theta|u_j) = -\frac{\partial^2}{\partial\theta^2}\log L(\theta|u_j) \quad (16)$$

The expected value of the observed FI in Equation 16 is the theoretical FI in Equation 12. The observed TIF and observed SEM are calculated in a similar way as their theoretical counterparts,

$$I(\theta|\boldsymbol{u}) = \sum_{j=1}^{J} I_j\left(\theta|u_j\right). \tag{17}$$

$$\mathrm{SEM}(\theta|\boldsymbol{u}) = \frac{1}{\sqrt{I(\theta|\boldsymbol{u})}} \tag{18}$$

Theoretical FI is typically used for item selection and test development, when no responses have been observed. Since the observed FI considers the response pattern and hence whether the observed responses fit the IRT model, the observed SEM is a better indicator of the uncertainty about the θ estimate than the theoretical SEM. For example, the observed SEM is often used as a stopping criterion of a computerized adaptive test (Weiss, 2011).

Chapter 3. Item Selection and Stopping Rules for Two-Category Classification Testing

Notation

Say we have a test that classifies examinees into two categories: mastery and non-mastery. Let θ_c be a pre-determined cutoff score such that for examinees whose $\theta \geq \theta_c$, the correct classification is mastery, while for examinees whose $\theta < \theta_c$, the correct classification is non-mastery. Let D be the classification decision made by the test, where $D = m$ when the decision is mastery, and $D = n$ when the decision is non-mastery. A false negative occurs when $\theta \geq \theta_c$ but $D = n$, whereas a false positive occurs when $\theta < \theta_c$ but $D = m$. Let T be the maximum test length, and $\boldsymbol{u}_t = (u_1, \ldots, u_t)$ be the response pattern of an examinee to the first t items of a test, where $t \leq T$.

Item Selection Methods

Maximum Fisher Information

Let S be the set of available items. Sequential tests that aim to maximize measurement precision often select the item that maximizes the Fisher information at $\hat{\theta}$,

$$\max_{j \in S} I_j(\hat{\theta}) \tag{19}$$

In CCT, one might also consider selecting the item that maximize the Fisher information at the cutoff,

$$\max_{j \in S} I_j(\theta_c) \tag{20}$$

Kullback-Leibler Divergence

The KL divergence measures the difference between two probability distributions, f and g, over the same variable x,

$$\mathrm{KL}(f||g) = E_f\left(\log\left[\frac{f(x)}{g(x)}\right]\right) = \int_{-\infty}^{\infty} f(x)\log\left[\frac{f(x)}{g(x)}\right]dx \quad (21)$$

where $f||g$ indicates the divergence of f from g. Eggen (1999) suggested maximizing the KL divergence between IRFs at two θ values around the cutoff,

$$\mathrm{KL}_j(\theta_c - \zeta||\theta_c + \zeta) = P_j(\theta_c - \zeta)\log\frac{P_j(\theta_c - \zeta)}{P_j(\theta_c + \zeta)} + Q_j(\theta_c - \zeta)\log\frac{Q_j(\theta_c - \zeta)}{Q_j(\theta_c + \zeta)}, \quad (22)$$

where ζ is a small constant. Most research used the width of the indifference region δ as ζ (e.g., Lau & Wang, 1999; Lin & Spray, 2000), but alternatives have been proposed. (Eggen, 1999).

A drawback of KL divergence is its lack of symmetry, which means $KL_j(\theta_1||\theta_2) \neq KL_j(\theta_2||\theta_1)$ in general. The methods described below do not suffer from this problem.

Weighted Log Odds Ratio

Lin and Spray (2000) proposed the use of a weighted log-odds ratio WLOR,

$$\mathrm{WLOR}_j(\theta_c + \delta||\theta_c - \delta) = \mathrm{E}(u_j = 1)\log\frac{P_j(\theta_c + \delta)}{P_j(\theta_c - \delta)} + \left[1 - \mathrm{E}(u_j = 1)\right]\log\frac{Q_j(\theta_c + \delta)}{Q_j(\theta_c - \delta)} \quad (23)$$

where $\mathrm{E}(u_j = 1)$ is the classical difficulty of item j and can be calculated by integrating the probability of response for θ weighted by the density of θ across the examinee distribution (Lin & Spray, 2000). The rationale for using this value to select items within the SPRT framework is that we are searching for items that will cause the likelihood ratio in Equation 23 to cross the decision boundaries most quickly (Lin & Spray, 2000). Thus, items with a greater value of the weighted log-odds ratio should be selected earlier,

$$\max_{j \in S} \mathrm{LO}_j(\theta_c + \delta||\theta_c - \delta)\,. \quad (24)$$

Expected Log Likelihood Ratio

Nydick (2013) replaced $\mathrm{E}(u_j = 1)$ in WLOR with $P_j(\hat{\theta})$, the probability of correct response given the current $\hat{\theta}$, and termed it the expected likelihood ratio (ELR):

$$\mathrm{ELR}_j(\hat{\theta}) = P_j(\hat{\theta}) \log \frac{P_j(\theta_c + \delta)}{P_j(\theta_c - \delta)} + Q_j(\hat{\theta}) \log \frac{Q_j(\theta_c + \delta)}{Q_j(\theta_c - \delta)}. \tag{25}$$

If $\hat{\theta} < \theta_c$, then the next item should be chosen to maximize Equation 25, whereas if $\hat{\theta} \geq \theta_c$, then the next item should be chosen to minimize Equation 25. Hu and Shih (2021) modified ELR to make it more suitable for GLR: If $\hat{\theta} > \theta_c + \delta$, select the item that

$$\max_j \mathrm{ELR}_j(\hat{\theta}) = P_j(\hat{\theta}) \log \frac{P_j(\hat{\theta})}{P_j(\theta_c - \delta)} + Q_j(\hat{\theta}) \log \frac{Q_j(\hat{\theta})}{Q_j(\theta_c - \delta)}. \tag{26}$$

If $\hat{\theta} < \theta_c - \delta$, select the item that

$$\max_j \mathrm{ELR}_j(\hat{\theta}) = P_j(\hat{\theta}) \log \frac{P_j(\theta_c + \delta)}{P_j(\hat{\theta})} + Q_j(\hat{\theta}) \log \frac{Q_j(\theta_c + \delta)}{Q_j(\hat{\theta})}. \tag{27}$$

Stopping Rules

Ability Confidence Interval

In the ability confidence interval (ACI; Thompson, 2009a), a confidence interval (CI) of $\hat{\theta}^{\mathrm{MLE}}$ is constructed after every item. Because $\hat{\theta}^{\mathrm{MLE}}$ is asymptotically normal for a single examinee across a sequence of items (Chang & Stout, 1993; Chang & Ying, 2009), the asymptotic $(1 - \alpha) \times 100.0\%$ CI for θ is given by (Finkelman, 2010; Sie et al., 2015)

$$\hat{\theta}^{\mathrm{MLE}} \pm z_{1-\alpha/2} \cdot \mathrm{SEM}(\hat{\theta}^{\mathrm{MLE}}) \tag{28}$$

where $z_{1-\alpha/2}$ denotes the $1 - \alpha/2$ quantile of a standard normal random variable, and $\mathrm{SEM}(\hat{\theta}^{\mathrm{MLE}})$ is defined in Equation 18. If the upper endpoint of the CI falls below θ_c, classify the examinee as non-mastery. If the lower endpoint is above θ_c, classify the

examinee as mastery. If θ_c is inside the CI, administer another item. Under normal circumstances, as more items are administered, the accumulated Fisher's information for $\hat{\theta}$ will increase, and the width of the CI will decrease. By selecting items strategically, the CI width can be shortened quickly, resulting in a shorter test. Item selection methods are discussed in a later section.

If a Bayesian method is used to estimate θ, one can use the variance of the posterior distribution to construct the credible interval instead, and the method is sometimes known as sequential Bayes in the literature (Kingsbury & Weiss, 1983; Spray & Reckase, 1994; Spray & Reckase, 1996).

Although most research uses the observed SEM to construct the CI (Ren et al., 2022), it was found that using the theoretical SEM results in a shorter test with almost no difference in classification accuracy (Thompson, 2011). ACI tends to require a longer test length than the method discussed below (Eggen & Straetmans, 2000; Spray & Reckase, 1996; Thompson, 2009a; Thompson, 2011).

Sequential Probability Ratio Test

The sequential probability ratio test (SPRT) was developed by Wald (1947) initially for quality control purposes. Ferguson (1969) was the first to apply SPRT for a classification testing scenario under a classical test theory framework. Reckase (1983) later applied it with IRT.

SPRT requires a pair of simple hypotheses. So instead of testing $H_0: \theta < \theta_c$ against $H_1: \theta \geq \theta_c$, we recast these hypotheses into

$$H_0: \theta = \theta_c - \delta$$

$$H_1: \theta = \theta_c + \delta$$

where δ is a small constant. After the administration of every item, compute the ratio between the likelihood of $\boldsymbol{u}_t$ evaluated at $\theta_c + \delta$ and the same likelihood evaluated at $\theta_c - \delta$,

$$\lambda^{\text{SPRT}} = \frac{L(\theta_c + \delta|\boldsymbol{u}_t)}{L(\theta_c - \delta|\boldsymbol{u}_t)} \tag{29}$$

Then, for some constants A and B, where $0 < A < B < \infty$, terminate the test and accept H_0 if $\log \lambda^{\text{SPRT}} \leq \log A$, terminate the test and accept H_1 if $\log \lambda^{\text{SPRT}} \geq \log B$; otherwise, administer another item. A false negative rate of β and a false positive rate of α can be approximately achieved by using (Wald, 1947)

$$A = \frac{\alpha}{1 - \beta} \tag{30}$$

$$B = \frac{1 - \alpha}{\beta} \tag{31}$$

The interval $(\theta_c - \delta, \theta_c + \delta)$ is called the indifference region. It accounts for the uncertainty of the decisions regarding examinees with θ close to the cutoff (Eggen, 1999). The wider the indifference region, the earlier the test will be terminated. This is because IRFs are strictly increasing in θ, a large value of δ will lead to a larger difference between $P_j(\theta_c + \delta)$ and $P_j(\theta_c - \delta)$ if the item was answered correctly, or $Q_j(\theta_c + \delta)$ and $Q_j(\theta_c - \delta)$ if the item was answered incorrectly (Thompson, 2011). This in turn allows $\log \lambda^{\text{SPRT}}$ to cross the decision boundaries $\log A$ and $\log B$ with fewer items. However, a wider indifference region will also lead to a decrease in classification accuracy (Huebner & Fina, 2015; Thompson, 2011; van Groen et al., 2014). Therefore, it is recommended not to specify the value of δ arbitrarily (Thompson, 2009b). Rather, one should experiment with a range of δ in a simulation to find the δ value that produces the

shortest test lengths while still maintaining the desired level of classification accuracy (Thompson, 2009b).

The original SPRT does not bound the number of items that can be administered. However, in practice, the test cannot continue indefinitely. When the test reaches the maximum length, a forced classification has to be made. This is sometimes known as the truncated sequential probability ratio test (TSPRT) in the literature (e.g., Bartroff et al. 2008; Finkelman, 2008). When the maximum test length is reached, Spray and Reckase (1996) suggested to accept H_0 if $\log \lambda^{\text{SPRT}} < F$, where

$$= \frac{\log A + \log B}{2} \tag{32}$$

and accept H_1 otherwise. Note that $F = 0$ when $\alpha = \beta$.

Generalized Likelihood Ratio Test

Bartroff et al. (2008) and Thompson (2009b) independently suggested re-formulating the classification problem as two composite hypotheses

$$H_0: \theta \leq \theta_c - \delta$$

$$H_1: \theta \geq \theta_c + \delta$$

and use the generalized likelihood ratio (GLR),

$$\lambda^{\text{GLR}} = \frac{\sup_{\theta \geq \theta_c + \delta} L(\theta | \boldsymbol{u}_t)}{\sup_{\theta \leq \theta_c - \delta} L(\theta | \boldsymbol{u}_t)} \tag{33}$$

It was argued that this conceptually matches the goal of CCT more closely (Thompson, 2011; Wang et al., 2021). When $\hat{\theta}^{\text{MLE}}$ is outside of the indifference region, and when the likelihood function of $\boldsymbol{u}_t$ is unimodal (which is true under normal circumstances), either the θ that satisfied the constraint in the numerator or the θ that satisfied the constraint in the denominator is $\hat{\theta}^{\text{MLE}}$. The other is an endpoint of the indifference region on the

opposite side of $\hat{\theta}^{\text{MLE}}$ (Wang et al., 2022). When $\hat{\theta}^{\text{MLE}}$ is inside the indifference region, the GLR reduces to SPRT (Wang et al., 2022). Therefore, Equation 33 can be re-expressed as

$$\lambda^{\text{GLR}} = \begin{cases} \dfrac{L\left(\hat{\theta}^{\text{MLE}} \middle| \boldsymbol{u}_t\right)}{L(\theta_c - \delta | \boldsymbol{u}_t)} \text{ if } \hat{\theta}^{\text{MLE}} > \theta_c + \delta \\ \dfrac{L(\theta_c + \delta | \boldsymbol{u}_t)}{L\left(\hat{\theta}^{\text{MLE}} \middle| \boldsymbol{u}_t\right)} \text{ if } \hat{\theta}^{\text{MLE}} < \theta_c - \delta \\ \dfrac{L(\theta_c + \delta | \boldsymbol{u}_t)}{L(\theta_c - \delta | \boldsymbol{u}_t)} \text{ otherwise.} \end{cases} \tag{34}$$

Intuitively, GLR compares $\hat{\theta}^{\text{MLE}}$ to the most likely value of the composite hypothesis to which $\hat{\theta}^{\text{MLE}}$ does not belong (Nydick, 2013). Contrary to Bartroff et al. (2008), who proposed to use a simulation and a rather complicated numerical method to determine A, B and F, Thompson (2009b) suggested that Equations 30, 31 and 32 can still be used. By using $\hat{\theta}^{\text{MLE}}$, GLR accounts better for observed responses than does simply using the two fixed points. It has been demonstrated that GLR produces shorter tests than SPRT without sacrificing classification accuracy (Huebner & Fina, 2015; Thompson, 2009b, 2011).

Stochastic Curtailment

The concept of stochastic curtailment (SC) originates from group sequential clinical trials (Lan et al., 1982). It is sometimes desirable to stop a clinical trial as soon as the decision is inevitable to prevent exposing human subjects to additional potential risks by continuing the trial. Finkelman (2008) adopted this idea and proposed to terminate the test when the tentative classification decision at the current stage of the test is unlikely to change, had the test continued. Stopping rules described in the previous sub-sections

consider the amount of information obtained from items already administered; SC considers the amount of information expected to gain from future items.

Let D_t be the classification decision made by SPRT after t items, where $D_t = m$ (mastery) if $\log \lambda_t^{\mathrm{SPRT}} < F$ and $D_t = n$ (non-mastery) otherwise. Also, denote $P_\theta(D_T = c|\lambda_t^{\mathrm{SPRT}})$ as the conditional probability that SPRT will ultimately classify the examinee as category $c = \{m, n\}$, given the likelihood ratio λ_t^{SPRT} at the current stage, with the expected responses to future items evaluated at θ. Finkelman (2008) followed Lan et al. (1982) in requiring that $P_\theta(D_T = c|\lambda_t^{\mathrm{SPRT}})$ be adequately high at both $\theta = \theta_c - \delta$ and $\theta = \theta_c + \delta$. This choice of θ will be referred to as the standard formulation of stochastic curtailment (SC-Standard). Under usual conditions, $P_\theta(D_T = m|\lambda_t^{\mathrm{SPRT}})$ increases, as θ increases. Since $\theta_c + \delta > \theta_c - \delta$, it follows that $P_{\theta_c+\delta}(D_T = m|\lambda_t^{\mathrm{SPRT}}) > P_{\theta_c-\delta}(D_T = m|\lambda_t^{\mathrm{SPRT}})$. Conversely, $P_\theta(D_T = n|\lambda_t^{\mathrm{SPRT}})$ decreases, as θ increases. Following a similar line of argument, $P_{\theta_c-\delta}(D_T = n|\lambda_t^{\mathrm{SPRT}}) > P_{\theta_c+\delta}(D_T = n|\lambda_t^{\mathrm{SPRT}})$. Putting it all together, SC-Standard can be reduced to the following: classify an examinee as a master if

$$\log \lambda_t^{\mathrm{SPRT}} \geq F \text{ and } P_{\theta_c-\delta}\left(D_T = m\middle|\lambda_t^{\mathrm{SPRT}}\right) \geq \gamma \tag{35}$$

and classify an examinee as a non-master if

$$\log \lambda_t^{\mathrm{SPRT}} < F \text{ and } P_{\theta_c+\delta}\left(D_T = n\middle|\lambda_t^{\mathrm{SPRT}}\right) \geq \gamma \tag{36}$$

where $0 \leq \gamma \leq 1$.

It is computationally burdensome to evaluate these probabilities exactly, especially in the early stages of the test, as this requires considering all possible response patterns to the remaining items. Hence, Finkelman (2008) suggested to use a central limit theorem (CLT) approximation. For example,

$$P_\theta\left(D_T = n \middle| \lambda_t^{\text{SPRT}}\right) = P_\theta\left(\log \lambda_T^{\text{SPRT}} < F \middle| \lambda_t^{\text{SPRT}}\right) \approx \Phi\left(\frac{F - \mathrm{E}_\theta\left(\log \lambda_T^{\text{SPRT}} \middle| \lambda_t^{\text{SPRT}}\right)}{\left[\mathrm{Var}_\theta\left(\log \lambda_T^{\text{SPRT}} \middle| \lambda_t^{\text{SPRT}}\right)\right]^{1/2}}\right), \quad (37)$$

where

$$\mathrm{E}_\theta\left(\log \lambda_T^{\text{SPRT}} \middle| \lambda_t^{\text{SPRT}}\right) = \log \lambda_t^{\text{SPRT}} + \sum_{t'=t+1}^{T}\left[P_{t'}(\theta) \log \frac{P_{t'}(\theta_c+\delta)}{P_{t'}(\theta_c-\delta)} + Q_{t'}(\theta) \log \frac{Q_{t'}(\theta_c+\delta)}{Q_{t'}(\theta_c-\delta)}\right] \quad (38)$$

$$\mathrm{Var}_\theta\left(\log \lambda_T^{\text{SPRT}} \middle| \lambda_t^{\text{SPRT}}\right) = \sum_{t'=t+1}^{T}\left\{\mathrm{E}_\theta\left[\left(\log \frac{P_{t'}(\theta_c+\delta)}{P_{t'}(\theta_c-\delta)}\right)^2\right] - \left[\mathrm{E}_\theta\left(\log \frac{P_{t'}(\theta_c+\delta)}{P_{t'}(\theta_c-\delta)}\right)\right]^2\right\} \quad (39)$$

are the conditional expectation and conditional variance of the final log likelihood ratio λ_T^{SPRT} given the current log likelihood ratio λ_t^{SPRT}. Obviously,

$$P_\theta\left(D_T = m \middle| \lambda_t^{\text{SPRT}}\right) = 1 - P_\theta\left(D_T = n \middle| \lambda_t^{\text{SPRT}}\right). \quad (40)$$

The summations in Equations 38 and 39 require the identification of future items. It is not necessary to determine the exact set of future items; the future items only need to be representative enough to provide a good approximation (Finkelman, 2008). For example, if the item selection method is maximum Fisher information at $\hat{\theta}^{\text{MLE}}$, then the set might include items that are maximally informative at the current $\hat{\theta}^{\text{MLE}}$. Huebner and Finkelman (2016) showed that the CLT approximation in Equation 37 works well early in the test when the number of remaining items is large. If test efficiency is a top priority, they recommended computing the exact probabilities when there are five to eight items remaining in the test.

Finkelman (2010), realizing that using information about $\hat{\theta}$ could potentially provide a better estimate of $P_\theta(D_T = c|\lambda_t)$ than the endpoints of the indifference region, proposed three variations on stochastic curtailment. The first variation simply uses $\theta = \hat{\theta}$ in Equations 38 and 39. This is known as the MLE formulation (SC-MLE). To take into consideration the uncertainty of the estimation of θ, Finkelman (2010) also suggested

using the endpoints of the CI of $\hat{\theta}$ (SC-CI). Denote the lower and upper endpoints of the CI for $\hat{\theta}$ as $\hat{\theta}_l$ and $\hat{\theta}_u$ respectively. To be more conservative, Finkelman (2010) recommended using the endpoint that results in a smaller value of the probabilities in Equation 35 and 36. That is, using $\theta = \hat{\theta}_l$ when $D_t = m$, and use $\theta = \hat{\theta}_u$ when $D_t = n$. Hence, SC-CI will be more conservative than SC-MLE. The uncertainty about $\hat{\theta}$ can also be quantified through a Bayesian approach (SC-Bayes). In particular, we can integrate $P_\theta(D_T = n|\lambda_t^{\text{SPRT}})$ over the posterior distribution of θ after t items, $\pi(\theta|\boldsymbol{u}_t)$ defined in Equation 9, to obtain the predictive posterior distribution,

$$P_{\pi(\theta|\boldsymbol{u}_t)}\left(D_T = n\middle|\lambda_t^{\text{SPRT}}\right) = \int_\theta P_\theta\left(D_T = n\middle|\lambda_t^{\text{SPRT}}\right)\pi(\theta|\boldsymbol{u}_t)d\theta \tag{41}$$

This probability can be used in place of $P_\theta(D_T = n|\lambda_t^{\text{SPRT}})$ in Equation 36. Similarly, $P_{\pi(\theta|\boldsymbol{u}_t)}(D_T = m|\lambda_t^{\text{SPRT}}) = 1 - P_{\pi(\theta|\boldsymbol{u}_t)}(D_T = n|\lambda_t^{\text{SPRT}})$ can be used in place of $P_\theta(D_T = m|\lambda_t^{\text{SPRT}})$ in Equation 35.

Finkelman (2010) recommended not to use these new variations during the early stages of the test, when $\hat{\theta}$ is imprecise and unstable. Rather, SC-Standard should be used, until either the standard error of $\hat{\theta}$ is below a given threshold, or a certain number of items have been administered, to avoid a premature termination of the test. All three variations were able to terminate the test earlier than SC-Standard, without loss in classification accuracy (Finkelman, 2010).

Chapter 4. Item Selection and Stopping Rules for Multi-Category Classification Testing

Notation

Assume that the goal of a test is to classify examinees into one of $C + 1$ categories, requiring C cutoffs, $\theta_1 < \ldots < \theta_c < \ldots < \theta_C$.

Item Selection Methods

Maximum Fisher Information

Just as in two-category CCT, it is possible to select the item that maximizes the Fisher information at $\hat{\theta}$. However, many other variations have been proposed for multi-category classification scenarios because there is more than one cutoff. Denote the nearest cutoff to $\hat{\theta}$ as $\theta_{c^\star}$ such that $c^\star \equiv \underset{c\in\{1,\ldots,C\}}{\arg\min}\left|\theta_c - \hat{\theta}\right|$. We can select the item that maximizes information at the nearest cutoff (Eggen, 2009; Eggen & Straetmans, 2000; Thompson, 2007; Wouda & Eggen, 2009),

$$\max_{j\in S} I_j(\theta_{c^\star}) \tag{42}$$

Veerkmap and Berger (1997) proposed selecting the item with the largest weighted Fisher information,

$$\max_{j\in S} \int_{\Theta} w_\theta \, I_j(\theta)\mathrm{d}\theta \tag{43}$$

where Θ can be a set of discrete points or an interval of θ values, and w_θ is the weight for a given θ. van Groen et al. (2014) suggested to use $\Theta = \{\theta_1, \ldots, \theta_C, \hat{\theta}\}$ and $w_\theta = 1$. That is, to select items that provide the maximum sum of information at each cutoff θ_c and $\hat{\theta}$,

$$\max_{j\in S} \sum_{\theta\in\{\theta_1,\ldots,\theta_C,\hat{\theta}\}} I_j(\theta) \tag{44}$$

To reduce the impact of cutoffs that are far away from $\hat{\theta}$, Weissman (2007) proposed to use $\Theta = \{\theta_1, \ldots, \theta_C\}$ and $w_\theta = P(\theta_c|\boldsymbol{u})$, the posterior density of θ,

$$\max_{j \in S} \sum_{c=1}^{C} P(\theta_c|\boldsymbol{u}) I_j(\theta_c) \tag{45}$$

Similarly, Wang et al. (2021) weighted the information of all cutoffs by the likelihood function of θ,

$$\max_{j \in S} \sum_{c=1}^{C} L(\theta_c|\boldsymbol{u}) I_j(\theta_c) \tag{46}$$

Kullback-Leibler Divergence

For a three-category test, Eggen (1999) proposed two alternatives. The first is to select the item with the maximum KL divergence around the cutoff point closest to $\hat{\theta}$.

$$\max_{j \in S} \mathrm{KL}_j(\hat{\theta}||\theta_{c^\star}) \tag{47}$$

This approach can obviously be generalized to tests with more than three categories. The second is as follows: when none of the pairs of hypotheses has led to a decision, items are chosen with maximum KL divergence between the two cutoffs θ_1 and θ_2, but if one of the pairs of hypotheses has led to a decision while the other has not, the items selected will have maximum KL divergence around the cutting point corresponding to the undecided test. That is, an item is selected for which

$$\text{If } \frac{L(\theta_1 + \delta|\boldsymbol{u}_t)}{L(\theta_1 - \delta|\boldsymbol{u}_t)} \geq B, \max_{j \in S} \mathrm{KL}_j(\theta_2 + \delta||\theta_2 - \delta) \tag{48}$$

$$\text{If } \frac{L(\theta_2 + \delta|\boldsymbol{u}_t)}{L(\theta_2 - \delta|\boldsymbol{u}_t)} \leq A, \max_{j \in S} \mathrm{KL}_j(\theta_1 + \delta||\theta_1 - \delta) \tag{49}$$

$$\text{Otherwise,} \quad \max_{j \in S} \mathrm{KL}_j(\theta_2||\theta_1) \tag{50}$$

A narrower interval can be substituted for Equation 50:

$$\max_{j \in S} \mathrm{KL}_j(\theta_2 - \delta || \theta_1 + \delta) \tag{51}$$

The second alternative has a better theoretical rationale, as it takes all cutoffs into account. However, simulation results showed that they resulted in very little difference in terms of average test length and classification accuracy (Eggen, 1999).

Mutual Information

Mutual information (MI; Shannon & Weaver, 1949) is a measure of the mutual dependence between the two variables X and Y. Specifically, it quantifies the amount of information obtained about X by observing Y, and vice versa.

$$\mathrm{MI}(X, Y) = \int_{x \in X} \int_{y \in Y} f(x, y) \log \frac{f(x, y)}{f(x) f(y)} \tag{52}$$

Mutual information can be viewed as the KL divergence from the product of the marginal distributions, $f(x)f(y)$, of the joint distribution, $f(x, y)$,

$$\mathrm{MI}(X, Y) = \mathrm{KL}(f(x, y) || f(x) f(y)) \tag{53}$$

When X and Y are independent, $f(x, y) = f(x)f(y)$ and $\mathrm{MI}(X, Y) = 0$. Otherwise, the joint distribution would always provide information in addition to the marginal information of the two distributions. Therefore, MI is always non-negative.

MI was first used as an item selection criterion by Weissman (2004; 2007): set X as the item response to an item j, and Y as a discrete random variable of θ values. Then, one chooses items that satisfy the following,

$$\max_{j \in S} \mathrm{MI}_j(u_j | \Theta) = \sum_{u_j=0}^{1} \sum_{\theta \in \Theta} P(u_j, \theta) \log \frac{P(u_j, \theta)}{P(u_j)\pi(\theta)} = \sum_{u_j=0}^{1} \sum_{\theta \in \Theta} P(u_j | \theta) \pi(\theta) \log \frac{P(u_j | \theta)}{P(u_j)}. \tag{54}$$

where $\Theta = \{\theta_1 - \delta, \theta_1 + \delta, \ldots, \theta_C - \delta, \theta_C + \delta\}$ is the set of the endpoints of all indifference regions.

Chapter 5. Stopping Rules

Ability Confidence Interval

The extension of ACI is very straightforward. One simply needs to continue administering items until the CI for $\hat{\theta}$ does not include any cutoff. If the maximum test length is reached, classify the examinee into the category in which $\hat{\theta}$ lies.

Multi-Hypothesis Sequential Probability Ratio Test

The multi-hypothesis sequential probability ratio test (mSPRT; Sobel & Wald, 1949) formulates two hypotheses around each cutoff θ_c,

$$H_{c0}: \theta \leq \theta_c - \delta$$

$$H_{c1}: \theta \geq \theta_c + \delta$$

and uses the following test statistic,

$$\lambda_c^{\text{mSPRT}} = \frac{L(\theta_c + \delta | \boldsymbol{u}_t)}{L(\theta_c - \delta | \boldsymbol{u}_t)} \tag{55}$$

where $c = 1, \ldots, C$.

Similar to SPRT, H_{c0} is accepted if $\log \lambda_c \leq \log A$ and H_{c1} is accepted if $\log \lambda_c \geq \log B$. If H_{10} is accepted, classify the examinee into category 1. If H_{C1} is accepted, classify the person into category $C + 1$. For $1 \leq c < C$, if H_{c1} and $H_{(c+1)0}$ are both accepted, classify the examinee into category $c + 1$. If none of the above is satisfied, administer another item. If the maximum test length is reached, classify the examinee into category $\tilde{c}$ where $\hat{\theta}^{MLE} \in \Theta_{c'}$, where $c' = 1, \ldots, C + 1$. The half width of the indifference region δ should be chosen such that the indifference regions around different cutoffs are not overlapped (van Groen et al., 2014; Wang et al., 2021).

There has been some concern that the Sobel and Wald approach might lead to conflicting conclusions when more than three hypotheses are considered (e.g., Ghosh, 1970). To solve this problem, Armitage (1950) suggested comparing all possible pairs of categories. That is, for each pair (m, n) where $m < n \in \{1, \ldots, C + 1\}$, consider two composite hypotheses indicating that the examinee belongs to category m and n are, respectively

$$H_m: \theta \leq \theta_m - \delta$$

$$H_n: \theta \geq \theta_{n-1} + \delta$$

Then, use the following test statistic

$$\lambda_{m,n}^{\text{mSPRT}} = \frac{L(\theta_{n-1} + \delta | \boldsymbol{u}_t)}{L(\theta_m - \delta | \boldsymbol{u}_t)} \tag{56}$$

However, Wang et al. (2021) demonstrated that this concern is unnecessary, as it very rarely happens when the likelihood function is symmetric around $\hat{\theta}$. Hence, only the Sobel and Wald approach is considered hereafter.

Multi-Category Generalized Likelihood Ratio Test

Wang et al. (2021) extended the GLR test from two-category classification to multi-category classification. Assume that there are *C* cutoffs such that the parameter space Θ is split into *C* + 1 disjoint sets, $\Theta_0 = \{\theta < \theta_1 - \delta\}$, $\Theta_1 = \{\theta_1 + \delta < \theta < \theta_1 - \delta\}$, …, $\Theta_C = \{\theta > \theta_C + \delta\}$. The multi-category GLR (mGLR) test forms a composite hypothesis,

$$H_{c'}: \theta \in \Theta_{c'}$$

Then, the test statistic is

$$\lambda_{c'}^{\text{mGLR}} = \frac{\sup_{\theta \in \Theta} L(\theta | \boldsymbol{u}_t)}{\sup_{\theta \in \Theta_{c'}} L(\theta | \boldsymbol{u}_t)} \tag{57}$$

The numerator is the unconstrained maximum likelihood, whereas the denominator is the maximum likelihood under the constraint that θ belongs to the set $\Theta_{c'}$.

If the current $\hat{\theta}^{\text{MLE}}$ is outside the indifference regions around all C cutoffs, the test stops when there exists some c' such that $\log \lambda_k^{\text{mGLR}} \geq \log(1/\alpha)$, for all $k \neq c'$. If the current $\hat{\theta}^{\text{MLE}}$ is inside the indifference region around θ_c, compute λ_c^{mSPRT} as in Equation 55 and follow the decision rule of mSPRT.

Multi-Category Stochastic Curtailment

Wang et al. (2021) also extended SC-Standard (Finkelman, 2008) to the multi-category classification scenario. The decision rule is as follows: classify the examinee into category 1 if

$$\log \lambda_{1t}^{\text{mSPRT}} < F \text{ and } P_{\theta_1+\delta}\left(\log \lambda_{1t}^{\text{mSPRT}} < F \middle| \lambda_{1t}^{\text{mSPRT}}\right) \geq \gamma, \tag{58}$$

classify the examinee into category $C+1$ if

$$\log \lambda_{Ct}^{\text{mSPRT}} > F \text{ and } P_{\theta_C-\delta}\left(\log \lambda_{Ct}^{\text{mSPRT}} > F \middle| \lambda_{Ct}^{\text{mSPRT}}\right) \geq \gamma, \tag{59}$$

classify the examinee into the category defined between θ_c and θ_{c+1} if,

$$\log \lambda_{ct}^{\text{mSPRT}} > F \text{ and } \log \lambda_{(c+1)t}^{\text{mSPRT}} < F \text{ and,}$$

$$P_{\theta_c-\delta}\left(\log \lambda_{ct}^{\text{mSPRT}} > F \middle| \lambda_{ct}^{\text{mSPRT}}\right) \geq \gamma \text{ and } P_{\theta_{c+1}+\delta}\left(\log \lambda_{(c+1)t}^{\text{mSPRT}} < F \middle| \lambda_{(c+1)t}^{\text{mSPRT}}\right) \geq \gamma, \tag{60}$$

otherwise, administer another item.

The three variations in Finkelman (2010) have not been studied under a multi-category classification testing scenario, although their extensions are very straightforward. For SC-MLE, simply evaluate the conditional probabilities in Equations 58, 59 and 60 at $\hat{\theta}^{\text{MLE}}$, instead of the endpoints of the indifference regions. For SC-Bayes, simply integrate those probabilities over the posterior distribution $\pi(\theta|\boldsymbol{u}_t)$ defined in Equation 9, instead of evaluating them at a single cutoff. For SC-CI, following

Finkelman's (2010) advice of using a more conservative value for computing the probabilities, classify the examinee into category 1 if

$$\log \lambda_{1t}^{\text{mSPRT}} < F \text{ and } P_{\widehat{\theta}_u}\left(\log \lambda_{1t}^{\text{mSPRT}} < F \middle| \lambda_{1t}^{\text{mSPRT}}\right) \geq \gamma, \quad (61)$$

classify the examinee into category $C + 1$ if

$$\log \lambda_{Ct}^{\text{mSPRT}} > F \text{ and } P_{\widehat{\theta}_l}\left(\log \lambda_{Ct}^{\text{mSPRT}} > F \middle| \lambda_{Ct}^{\text{mSPRT}}\right) \geq \gamma, \quad (62)$$

classify the examinee into the category defined between θ_c and θ_{c+1} if,

$$\log \lambda_{ct}^{\text{mSPRT}} > F \text{ and } \log \lambda_{(c+1)t}^{\text{mSPRT}} < F \text{ and,}$$

$$P_{\widehat{\theta}_l}\left(\log \lambda_{ct}^{\text{mSPRT}} > F \middle| \lambda_{ct}^{\text{mSPRT}}\right) \geq \gamma \text{ and } P_{\widehat{\theta}_u}\left(\log \lambda_{(c+1)t}^{\text{mSPRT}} < F \middle| \lambda_{(c+1)t}^{\text{mSPRT}}\right) < 1 - \gamma, \quad (63)$$

Chapter 6. Literature Review

Studies Comparing Item Selection Methods

Spray and Reckase (1994) compared different item selection methods under two stopping rules, ACI and SPRT, in a two-category CCT. For ACI, they compared maximizing FI at the cutoff score, at the examinee's true θ, and at the examinee's $\hat{\theta}$. For SPRT, they compared maximizing FI at the cutoff score and at the examinee's true θ. The simulation was repeated with three different cutoff scores, -0.5, 0.0 and 1.0. It was found that maximizing FI at the cutoff score resulted in the shortest average test length (ATL) for both stopping rules at most θ levels. However, it is difficult to judge whether their conclusion was valid without more context. First, they imposed a strong Bayesian prior of $\mathrm{N}(0,1)$ in the construction of the CI around $\hat{\theta}$, but did not provide any justification for their choice of prior. When the cutoff score was at $\theta = 1.0$, high θ examinees needed more items to overcome this strong prior. Meanwhile, for low θ examinees, one incorrect response was often enough to produce a CI that was entirely below the cutoff. The authors admitted that this might result in high misclassification rates, but they did not report the classification accuracy for any condition. Second, they did not report or manipulate the information function of the item bank. It is likely that their item bank did not provide enough information at the low ranges or high ranges of θ, and thus was not ideal to employ the item selection methods that are adaptive to the examinee's θ.

Eggen (1999) compared various FI-based and KL-based item selection methods in two- and three-category CCT using SPRT as the termination criterion. He varied α, β and the width of the indifference region. For two-category CCT, given the same condition,

the largest difference in ATL among the methods was no more than one item. For three-category CCT, the difference in ATL between the best FI-based method and the best KL-based method was also smaller than one item within a given condition. All methods had a similar level of classification accuracy. He recommended KL-based methods, as some FI-based methods require the computation of $\hat{\theta}$, whereas KL-based methods only compared the likelihoods at two fixed points. However, such computation is trivial with modern day's processing power, so his comment is no longer relevant.

Lau and Wang (1999) compared maximizing FI at the cutoff and KL divergence around the cutoff in a two-category CCT with polytomous items. The stopping rule was SPRT. Independent variables manipulated included item exposure control methods, location of the cutoff, item bank size, and width of the indifference region. Just as Eggen (1999), they found nearly identical results in the two item selection methods across all conditions.

Eggen and Straetmans (2000) compared ACI and SPRT in a three-category CCT with content balancing and item exposure control. For ACI, they compared maximizing FI at the nearest cutoff and at $\hat{\theta}$. They found that maximizing FI at $\hat{\theta}$ led to a shorter test and had the same classification accuracy as maximizing FI at the nearest cutoff, which is the opposite to what Spray and Reckase (1994) found for two-category CCT. They explained that the discrepancy might be due to the differences in the characteristics of the item bank used, the number of cutoffs, or the use of content balancing and item exposure control. For SPRT, the only selection method was maximum FI at the nearest cutoff, so no comparison can be made.

Lin and Spray (2000) compared FI at the cutoff, KL divergence around the cutoff and WLOR in a two-category CCT using SPRT as the stopping rule. They varied the item exposure control method, item bank size, and the width of the indifference region. They found that the three item selection methods yielded very small differences in the classification accuracy and ATL, regardless of the conditions imposed.

Weissman (2004, 2007) compared FI at $\hat{\theta}$, posterior weighted FI, and MI in a four-category CCT with SPRT as the stopping rule. Weissman calculated the percentage of cumulative correct classifications and the percentage of cumulative incorrect classifications out of all examinees after every item, both within and across the four categories. The ATL was also calculated after every item, based on the examinees for whom SPRT made a confident decision and terminated the test. Results showed that maximum MI item selection generally resulted in higher percentages of correct and incorrect classifications than the other two methods, especially in the early stages of the test. The ATL for the maximum MI method tended to be shorter in all stages of the test. The only exception was for the highest category, where no superiority of any method was found. The author believed that this unstable pattern might be explained by relatively small size of the sample in that category. However, the results of this study were difficult to interpret because of the exclusion of examinees for whom the SPRT could not make a decision. If the percentage of correct classification was calculated based on examinees who were classified by SPRT, instead of all examinees, the conclusion might have been different. Moreover, examinees who reached the maximum test length were excluded from the calculations, if SPRT had yet to classify them, possibly resulting in an unrealistic portrayal of ATL.

Thompson (2009a) hypothesized that item-selection methods can be generally classified as estimate-based or cutscore-based. He reasoned that maximizing FI at $\hat{\theta}$ is more suitable for ACI, as this can shorten the CI more quickly and hence facilitate ACI to make a decision. On the other hand, for SPRT, a decision is made more quickly when the likelihood ratio is maximized or minimized. Thus, one should choose a cutscore-based item selection method to facilitate this. In his simulation, item selection methods (maximum FI at $\hat{\theta}$ and maximum FI at cutoff), stopping rules (ACI and SPRT), item bank information functions (flat vs peaked), and item bank sizes (300 items vs 750 items) were crossed. The results showed that ACI did work better with FI at $\hat{\theta}$, and SPRT worked better with FI at cutoff. Although one might think a flat item bank is more appropriate for estimate-based selection, the study found that peaked item banks are more efficient for both stopping rules. Thompson (2009a) explained that it was likely due to the fact that examinees far from the cutoff did not need many items to be classified; the extra information in those regions was not necessary.

Wouda and Eggen (2009) compared three item selection methods in a three-category CCT, including maximizing FI at $\hat{\theta}$, the nearest cutoff, and the average of the two cutoffs. Two stopping rules, SPRT and SPRT with SC-Standard, were considered. It was found that maximizing FI at the average of the two cutoffs had the longest ATL, regardless of the stopping rule. It also had the highest classification accuracy, although the advantage was less than 1%. There was little difference between maximizing FI at $\hat{\theta}$ and at the nearest cutoff.

Lin (2011) compared maximizing FI at cut score, KLI, WLOR, and MI in a two-category CCT. Other variables manipulated included whether item exposure control and

content balancing were imposed, the θ distribution, and the width of the indifference region. It was found that WLOR worked best in all conditions, but the differences among the four item selection methods were washed out as more realistic constraints were imposed. The only stopping rule considered in this study was SPRT.

van Groen et al. (2014) proposed four item selection methods that could consider multiple cutoffs simultaneously such as maximizing the sum of FI at all cutoffs weighted by the reciprocal of the absolute distance between the cutoff and $\hat{\theta}$. They were compared with methods that only considered one point: maximum FI at $\hat{\theta}$, at the nearest cutoff, and at the middle of the nearest sets of cutoffs. Random selection was also considered as a baseline. Other variables manipulated included the number of cutoffs, the imposition of item exposure control and content balancing, and the width of the indifference region. SPRT was used as the stopping rule. It was found that the methods that only considered one point worked just as well as the four proposed methods that considered multiple points.

To summarize, no existing literature has compared item selection methods with GLR as the stopping rule. One can observe from Equation 34 that GLR requires the evaluation of the likelihood at both $\hat{\theta}$ and the cutoffs, so it is not clear that whether an estimate-based method or a cutscore-based method is more suitable for GLR. Thompson (2009a, 2011) believed that a cutscore-based method was more suitable, while Wang et al. (2021) suspected the opposite. Neither of them verified their hypothesis in their studies.

Studies on Item Calibration Errors

All studies discussed earlier designed their item banks by generating item parameters with values drawn directly from a specified distribution. In a realistic setting, item banks are created by administering a set of items to a calibration sample and then using the responses to estimate the item parameters. This estimation process inherently introduces error into the item bank. Studies that have examined the impact of such error are reviewed below.

Hambleton et al. (1993) were one of the first to study the effect of item calibration errors in the selection of test items during test development. They simulated items according to a 1PL IRT model, so that all items had a true discrimination equal to 1.0. Then, θs were generated from a standard normal distribution and item responses were generated from those θs. To highlight the fact that any variation in the item discrimination estimates was solely due to estimation, a 2PL model was fitted to the data. Finally, the best 25 items to provide the target test information function were selected using the item parameter estimates. They believed that two variables are important in determining the size of the impact of item estimation errors: (1) the sample size in item calibration and (2) the ratio of item bank size to test length. The first variable is important, because it is inversely related to the size of the item parameter estimation errors. The second variable is important, because the larger the item bank and the shorter the test, the more opportunity there is to capitalize on chance by selecting spuriously high-discriminating items. Their results showed that all the selected items had a higher discrimination than their true values, and therefore the test information was overestimated. These effects were indeed more evident when a smaller calibration sample

and a larger ratio of item bank size to test length were used. Hambleton et al. (1993) referred the problem of selecting items with overestimated information due to item calibration error as capitalization on chance.

Hambleton and Jones (1994) did a follow-up study to Hambleton et al. (1993). They used the item parameters from 80 items in the Graduate Management Admission Test (GMAT). The items in GMAT were chosen because they were fitted with a 3PL model, so the item calibration errors were expected to be larger than those associated with simpler models. Three variables were manipulated: calibration sample size, ratio of item bank size to test length, and whether content balancing was considered during item selection. Similar to Hambleton et al. (1993), item responses were generated for examinees with θs simulated from a standard normal distribution. A 3PL model was fitted to obtain item parameter estimates for the 80 items. It was again confirmed that a smaller calibration sample and a larger ratio of item bank size to test length led to an overestimation of test information. The magnitude of overestimation was greater when content balancing was imposed. However, the authors noted that no generalization of this finding should be made because the finding is specific to the content specification used in that study.

van der Linden and Glas (2000) examined the impact of capitalization on item calibration error in computerized adaptive testing (CAT). They simulated an item bank containing items calibrated with different sample sizes. It was found that items calibrated in the smaller samples had higher exposure rates for all θ levels. Also, the smaller the ratio of item bank size to test length, the stronger the effect. This effect was robust with respect to different item selection criteria (maximum information, minimum expected

posterior variance, and maximum expected posterior-weighted information) and different θ estimation methods (MLE and Bayesian). It was also found that the mean absolute error of the θ estimates tended to be larger when the ratio of item bank size to test length was larger. Incorporating an item exposure control method mitigated the problem, although the authors noted that the primary reason for incorporating exposure control in CAT should be test security, instead of the danger of capitalization on estimation errors.

Olea et al. (2012) also studied the effects of capitalization on chance in CAT. Their results showed that the estimation errors of the discrimination parameter for the items administered in the CAT were in general larger than those in a random test. For the difficulty parameter, the CAT condition did not show a larger systematic estimation error than the random test condition. The estimation errors of the pseudo-guessing parameters were larger in the items administered in the CAT for the lower θ levels. They also found that the proportion of items administered in CAT for which the discrimination parameter estimate exceeded its corresponding true value was larger for the central θ levels. This indicates that the effect of capitalization on chance are not the same for different levels of θ. The overestimated discrimination parameters in turn led to an overestimation of the precision of θ estimates by as much as 40% in some conditions. Finally, they compared two item exposure control methods, *b*-matching and progressive method. The *b*-matching method resulted in a greater reduction in the impact of capitalization on chance, but also had larger θ estimation errors.

Unlike van der Linden and Glas (2000), and Olea et al. (2012), who studied fixed-length CAT, Patton et al. (2013) examined the effects of capitalization on chance when a variable-length termination rule was used in a two-category CCT. Specifically, they

considered two testing formats. The first format terminated the test when the standard error of the θ estimate fell below a certain threshold. The classification decision was made by comparing the final θ estimate with a predetermined cutoff. The second format used the ACI stopping rule. Manipulated variables include the calibration sample size and the location of the cutoffs. Consistent with previous research on fixed-length CAT, the authors found that selecting items based on the maximum information criterion capitalized on item calibration errors, yielding spuriously high values of test information, especially when the calibration sample size was small. When the standard error rule was used as the test termination criterion, the tests were spuriously short. As the calibration sample size decreased, the recovery of θ estimates worsened. In contrast, the effect of calibration sample size on average test length for the ACI condition was quite small. The authors believed that this is because the ACI stopping rule depends on the location of the θ estimate, in addition to the size of the standard error. Regardless of the stopping rule, the effect of calibration sample size on classification accuracy was quite small when the cutoff was located at $\theta_c = 0.5$. However, a larger effect was observed when a more extreme cutoff was used ($\theta_c = 1.5$), possibly because fewer items were available. So far, this is the only study that examined the effects of capitalization on chance in CCT. It is, therefore, worthwhile to investigate whether the results in this study can be generalized to other stopping rules, because some stopping rules, such as GLR, also involve θ estimates and the standard errors.

Chapter 6: Simulation Study

Purposes

As stated in the introduction, this simulation aimed to compare SC-MLE and SC-CI with SC-Standard, examine which item selection methods work best for GLR in the context of multi-category CCT, and investigate the impact of item calibration errors.

Item Banks

Two item banks were generated, each consisting of 300 items. The first item bank had a flat information function, providing approximately equal information across a wide range of θ. The item parameters of this item bank were generated with $a \sim U[0.5, 1.5]$, $b \sim U[-3, 3]$ and $c \sim U[0, 0.25]$, following Finkelman (2010). This item bank can provide equiprecise measurement of θ and is presumably ideal for SC-MLE and SC-CI, which are both based on the estimate of θ. To investigate whether their superiority to SC-Standard still held in a more practical item bank, the second item bank was designed to concentrate information in a narrow range of θ. The item parameters of the peaked item bank were generated with $a \sim U[0.5, 1.5]$, $b \sim N(0, 1)$ and $c \sim U[0, 0.25]$. The first type will be referred to as a broad item bank, the second type as a peaked item bank.

Item Calibration Error

To examine the impacts of item calibration error, the simulation was conducted with the true item parameters as well as the estimated item parameters. A total of 1,000 θ values for the calibration sample were generated from $N(0, 1^2)$. This sample size was within the range of sample size recommended for the 3PL model (e.g., De Ayala, 2013; Hulin et al., 1982; Lord, 1968; Sahin & Anil, 2017). Item responses for the simulees were generated based on the true item parameters. Using the simulated responses, the item

parameter values were estimated using the R package *mirt* (Chalmers, 2012). The estimated item parameters were used for subsequent item selection, θ estimation, and stopping rules. Figure 1 presents the item bank information functions for the two simulated item banks.

Figure 1. Item Bank Information Functions

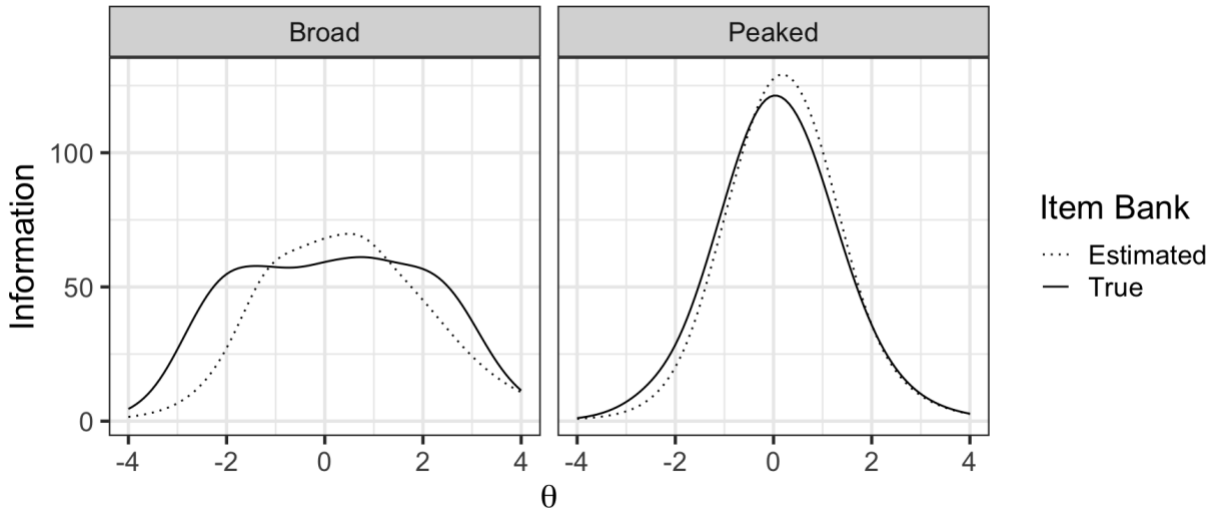

Table 1 presents the item parameter recovery statistics within each estimated item bank, including the bias, the root mean square error (RMSE), and the correlation between the true item parameters and the estimated item parameters. Overall, the peaked item bank had better item parameter recovery, so the information function of the estimated peaked item bank more closely resembled that of the true item bank (Figure 1). This is most likely due to the fact that the broad item bank contained more items with very low and very high difficulty parameters, but the responses were generated from a calibration sample whose θs were normally distributed.

Table 1. Item Parameter Recovery Statistics for Estimated Item Banks

	Bias			RMSE			Correlation		
Item Bank	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
Broad	0.00	-0.19	-0.09	0.40	0.46	0.21	0.71	0.97	0.24
Peaked	-0.09	-0.08	-0.03	0.28	0.24	0.12	0.87	0.97	0.51

Number of Cutoff and θ Distributions

This study used two sets of cutoffs. The first had two cutoff values: $(-0.44, 0.44)$, which correspond to the 33rd and 67th percentiles of a standard normal distribution, and was used by van Groen et al. (2014) and Wang et al. (2021). The second had three cutoff values: $(-1, 0, 1)$, which correspond to the 16th, 50th and 84th percentiles of a standard normal distribution, and was used by Spray (1993) and Wang et al. (2021).

Following Wang et al. (2021), for the two combinations of cutoffs, a sample of 1,000 θs was generated from the standard normal distribution, as well as two samples of 1,000 θs from a normal distribution with a smaller variance centered at different locations of the θ scale. Specifically, when the cutoffs were (–0.44, 0.44), 1,000 θs each were simulated from N(0, 1^2), N(–0.44, 0.2^2) and N(1, 0.2^2). Lastly, when the cutoffs were (–1, 0, 1), 1,000 θs each were simulated from N(0, 1), N(–0.5, 0.2^2) and N(–1.5, 0.2^2). The idea was to have one of the distributions centered at the cutoff, one centered at the middle of two cutoffs, and one located further away from all cutoffs. It was expected that when the mean of the generating distribution was close to the cutoff, the test length and classification error would increase.

Item Selection Methods

Thus far, no study has examined which item selection method works best with mGLR. Therefore, this study compared three item selection methods: maximum FI at $\hat{\theta}$

(FI-E), maximum FI at the nearest cutoff (FI-N), and maximum sum of FI at all cutoffs weighted by the likelihood function (FI-W). Based on Thompson's (2009a) classification, FI-E is an estimate-based method, while FI-N is a cutscore-based method. FI-W was used by Wang et al. (2021) and is a mix of both types.

Stopping Rules

Both mSPRT and mGLR were used with and without three SC methods (SC-Standard, SC-MLE, and SC-CI). For mSPRT and mGLR, $\alpha = \beta = 0.05$ (Finkelman, 2008, 2010; Lin, 2011). The probability threshold γ for all SC methods was fixed at 0.95, a value often used in previous studies (e.g., Finkelman, 2008, Wang et al., 2021). Finkelman (2010) recommended requiring a minimum test length, as SC-MLE and SC-CI all depend on information about $\hat{\theta}$ which is typically very imprecise during the early stages of the test. Thus, all SC methods were only applied after the 10th item. The maximum test length was fixed at 50 items.

Width of Indifference Regions

Van Groen et al. (2014) noted that "the size of the indifference region had little influence on accuracy but considerable influence on efficiency". Both Thompson (2011) as well as Huebner and Fina (2015), compared two levels of δ, 0.1 and 0.2, and found that $\delta = 0.2$ led to shorter test lengths with almost no sacrifice of accuracy. To verify that these observations were still valid in a multi-category classification scenario, the current study manipulated the half-width of indifference regions, δ, at four levels: 0.1, 0.2, 0.3 and 0.4, which are within the range of δ used in previous studies (Finkelman, 2008; Finkelman, 2010; Huebner & Fina, 2015; Lin, 2011; Thompson, 2011; van Groen et al., 2014; Wang et al., 2021). The indifference region for the first set of cutoffs would

overlap if $\delta \geq 0.5$ was used, which is against the recommendation by van Groen et al. (2014) and Wang et al. (2021).

Response Data Generation

To simulate the response to item j for an examinee with ability θ, a random number from $\mathrm{U}[0, 1]$ was first generated. If the random number was equal to or less than $P_j(\theta)$ defined in Equation 4, the simulee was said to answer the item correctly; incorrect otherwise.

The test for each simulee was first simulated to the maximum length using the item selection method in the given condition. Different stopping rules were then applied retroactively. This ensured a fair comparison: differences between condition would not be due to differences in response patterns. θ was estimated using MLE with a boundary of [-3, 3].

Summary

As described above, a variety of approaches to multi-category CCT were simulated. Ultimately, there were seven independent variables. They are summarized as follows:

1. Item bank information function: broad, peaked
2. Whether item calibration errors were introduced
3. Cutoff scores: (–0.44, 0.44), (–1, 0, 1)
4. θ distributions:
    a. For two-cutoff conditions, $\mathrm{N}(0, 1^2)$, $\mathrm{N}(–0.44, 0.2^2)$ and $\mathrm{N}(1, 0.2^2)$
    b. For three-cutoff conditions, $\mathrm{N}(0, 1^2)$, $\mathrm{N}(–0.5, 0.2^2)$ and $\mathrm{N}(–1.5, 0.2^2)$

5. Stopping rule: mSPRT, mGLR, mSPRT + SC-Standard, mSPRT + SC-MLE, mSPRT + SC-CI, mGLR + SC-Standard, mGLR + SC-MLE, mGLR + SC-CI
6. Item selection method: maximum FI at the nearest cutoff (FI-C), at $\hat{\theta}$ (FI-E), at all cutoffs weighted by the likelihood function (FI-W)
7. Width of indifference region: $\delta = 0.1, 0.2, 0.3, 0.4$

The average test length (ATL) and the percentage of correct classification (PCC) were evaluated. Results were analyzed to identify conditions that resulted in the best balance of ATL and PCC.

Chapter 7. Results

The ATL and PCC for all conditions described in Chapter 6 are displayed in Tables A-1 to A-16 in the Appendix. The marginal ATL and PCC for each level of each factor are presented in Table 2. Table 2 shows that the peaked item bank required an average of 2.4 fewer items than the broad item bank, while the PCC was 1.4% higher. Introducing item parameter estimation errors reduced the ATL from 22.9 items to 22.0 items, and slightly increased the PCC from 90.5% to 90.8%.

The θ distributions had a large impact on ATL and PCC. Table 2 shows that for the 2-cutoff conditions, when the mean of the θ distribution was far away from the cutoff point, i.e., $N(1, 0.2^2)$, the ATL was the shortest (15.1 items) and the PCC was the highest (98.2%). When the mean of the θ distribution was in the middle of two cutoffs, i.e., $N(0, 1^2)$, the ATL increased to 21.2 items, and the PCC dropped to 90.6%. The most challenging scenario was when the θ distribution was centered at a cutoff, i.e., $N(-0.44, 0.2^2)$. In this case, the ATL further increased to 28.5 items and the PCC went down to 77.5%. When there were three cutoffs, making a correct decision was more difficult. When the mean of the θ distribution was far away from a cutoff point, i.e., $N(1.5, 0.2^2)$, the ATL was the shortest (16.9 items) and the PCC was the highest (97.3%). When the mean of the θ distribution was in the middle of two cutoffs [i.e., $N(-0.5, 0.2^2)$], the ATL increased to 27.0 items, and the PCC dropped to 93.2%. Similar to the 2-cutoffs conditions, when the θ distribution was centered at a cutoff, i.e., $N(0, 1^2)$, the PCC was lowest.

All three item selection methods performed comparably in terms of ATL and PCC (Table 2). FI-W achieved the shortest ATL of 22.1 items, but it was only 0.7 items

shorter than the worst performing method, FI-E. The PCC of FI-C and FI-W was both 90.8%, which was only marginally better than the 90.3% of FI-E.

With regard to stopping rules, mGLR produced lower ATL than mSPRT (22.8 items vs 26.2 items). Among all SC methods, SC-MLE further reduced the ATL the most, to 18.6 items. Specifically, it reduced the ATL by 7.6 items when the primary stopping rule was mSPRT, and 4.2 items when the primary rule was mGLR. The difference between SC-CI and SC-Standard was less than one item, regardless of which primary stopping rule was used. The choice of stopping rules had very little impact on PCC. The mSPRT, mSPRT + SC-Standard and mSPRT + SC-CI tied for the highest PCC (91.0%), while mSPRT + SC-MLE and mGLR + SC-MLE tied for the lowest PCC (89.6%). Overall, mGLR + SC-MLE had the shortest ATL without affecting PCC.

The width of the indifference region had a noticeable effect on ATL and PCC. In particular, when δ increased from 0.1 to 0.4, the ATL decreased from 31.2 items to 15.4 items, although the PCC slightly decreased from 91.3% to 89.4%.

Table 2. Marginal ATL and PCC for Each Level of Each Factor

Condition	ATL	PCC
Item Bank Shape		
Broad	23.6	89.9
Peaked	21.2	91.3
Item Parameters		
True	22.9	90.5
Estimated	22.0	90.8
Cutoffs and θ Distributions		
(-0.44, 0.44)		
$N(0, 1^2)$	21.2	90.6
$N(-0.44, 0.2^2)$	28.5	77.5
$N(1, 0.2^2)$	15.1	98.2
(-1, 0, 1)		
$N(0, 1^2)$	25.8	86.9
$N(-0.5, 0.2^2)$	27.0	93.2
$N(1.5, 0.2^2)$	16.9	97.3
Item Selection Methods		
FI-C	22.3	90.8
FI-E	22.8	90.3
FI-W	22.1	90.8
Stopping Rules		
mSPRT	26.2	91.0
mGLR	22.8	90.9
mSPRT + SC-Standard	24.1	91.0
mSPRT + SC-MLE	19.8	89.6
mSPRT + SC-CI	23.7	91.0
mGLR + SC-Standard	22.1	90.9
mGLR + SC-MLE	18.6	89.6
mGLR + SC-CI	22.0	90.9
δ		
0.1	31.2	91.3
0.2	24.3	91.2
0.3	18.8	90.5
0.4	15.4	89.4

Table 3. Marginal ATL and PCC Conditional on Item Bank Shape and Item Selection Method

	ATL		PCC	
Item Selection Method	Broad	Peaked	Broad	Peaked
FI-C	23.4	21.2	90.1	91.4
FI-E	24.0	21.6	89.7	90.9
FI-W	23.2	20.9	90.0	91.5

Table 4. Marginal ATL and PCC Conditional on Item Bank Shape and Stopping Rule

	ATL		PCC	
Stopping Rule	Broad	Peaked	Broad	Peaked
mSPRT	27.8	24.7	90.3	91.6
mGLR	24.1	21.4	90.2	91.6
mSPRT + SC-Standard	25.2	23.0	90.3	91.6
mSPRT + SC-MLE	20.7	18.9	88.9	90.4
mSPRT + SC-CI	24.8	22.6	90.3	91.6
mGLR + SC-Standard	23.2	20.9	90.2	91.6
mGLR + SC-MLE	19.6	17.6	88.9	90.3
mGLR + SC-CI	23.2	20.9	90.3	90.2

Table 3 presents the marginal ATL and PCC conditional on item bank shape and item selection method. Table 4 presents the marginal ATL and PCC conditional on item bank shape and stopping rule. The results showed that the peaked item bank had shorter ATL (maximum 3.1 items) and slightly higher PCC after controlling for item selection method and stopping rule. These results may be somewhat unintuitive because a broad item bank might be more appropriate for item selection methods and stopping rules that evaluate information at the current $\hat{\theta}$. However, the current study found that the peaked

item bank resulted in more efficient and accurate tests for all item selection methods and stopping rules. This is likely due to the fact that examinees not near the cutoffs do not need many items to be classified. The extra information in those regions is not utilized in the broad item bank. On the other hand, the peaked item bank had more information near the cutoffs (Figure 1). Examinees with θ values near cutoffs are typically more difficult to classify, so the information in the peaked item bank is better utilized.

Table 5. Marginal ATL and PCC Conditional on Item Selection Method and Stopping Rule

	ATL			PCC		
Stopping Rule	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
mSPRT	26.0	27.1	25.7	91.1	90.8	91.1
mGLR	22.6	23.2	22.4	91.0	90.7	91.0
mSPRT + SC-Standard	24.0	24.7	23.6	91.1	90.8	91.1
mSPRT + SC-MLE	19.9	19.9	19.6	89.9	89.1	90.0
mSPRT + SC-CI	23.8	23.9	23.4	91.1	90.7	91.1
mGLR + SC-Standard	21.9	22.5	21.7	91.0	90.7	91.0
mGLR + SC-MLE	18.6	18.7	18.4	89.8	89.1	89.9
mGLR + SC-CI	21.9	22.5	21.7	91.0	90.7	91.0

Table 5 displays the marginal ATL and PCC conditional on item selection method and stopping rule. There was little interaction between item selection methods and stopping rules. After controlling for the stopping rules, FI-W still yielded the shortest ATL, followed by FI-C, and then FI-E. The largest difference between them (1.4 items) was observed when mSPRT was used as the stopping rule. The PCCs for all item selection methods were virtually the same, as the largest difference was no more than 1%.

Table 6. Marginal ATL and PCC Conditional on δ and Stopping Rule

Stopping Rule	ATL				PCC			
	δ = 0.1	δ = 0.2	δ = 0.3	δ = 0.4	δ = 0.1	δ = 0.2	δ = 0.3	δ = 0.4
mSPRT	42.5	27.2	19.6	15.7	91.7	91.7	90.9	89.7
mGLR	31.1	24.9	19.4	15.6	91.7	91.5	90.8	89.6
mSPRT + SC-Standard	34.5	26.6	19.5	15.7	91.7	91.7	90.9	88.8
mSPRT + SC-MLE	26.9	20.8	16.9	14.6	90.1	90.1	89.6	89.6
mSPRT + SC-CI	33.5	26.2	19.5	15.7	91.7	91.6	90.9	89.6
mGLR + SC-Standard	29.0	24.4	19.3	15.6	91.7	91.5	90.8	89.6
mGLR + SC-MLE	23.2	19.8	16.7	14.6	90.1	90.0	89.5	88.8
mGLR + SC-CI	28.9	24.4	19.3	15.6	91.7	91.5	90.8	89.6

Table 6 shows the marginal ATL and PCC conditional on δ and stopping rule. The results show that the differences between stopping rules—primarily in ATL— also became smaller as δ increased. For example, when $\delta = 0.1$, mGLR used an average of 11.4 fewer items than mSPRT. When $\delta = 0.4$, the difference was only 0.1 item. There were substantial differences between ATL for $\delta = 0.1$ versus $\delta = 0.4$: the largest difference was a reduction of 26.8 for SPRT and the minimum difference was 8.6 items for mGLR + SC-MLE. The largest reduction in PCC between the two δ conditions was 2.1 items. For all values of δ, mGLR + SC-MLE had the best balance between ATL and PCC.

Table 7. θ Recovery Statistics Conditional on Number of Cutoffs, θ Distribution, and True and Estimated Item Banks

Number of Cutoffs and θ Distribution	Bias		RMSE		Correlation	
	True	Estimated	True	Estimated	True	Estimated
(-0.44, 0.44)						
$N(0, 1^2)$	-0.02	-0.04	0.29	0.28	0.92	0.92
$N(-0.44, 0.2^2)$	0.05	0.05	0.11	0.10	0.62	0.61
$N(1, 0.2^2)$	-0.23	-0.22	0.42	0.45	0.41	0.45
(-1, 0, 1)						
$N(0, 1^2)$	-0.01	-0.02	0.14	0.12	0.95	0.95
$N(-0.5, 0.2^2)$	0.01	0.00	0.06	0.06	0.62	0.61
$N(1.5, 0.2^2)$	-0.19	-0.24	0.39	0.29	0.44	0.43

Table 7 presents the θ recovery statistics within true and estimated item banks as well as number of cutoffs and θ distributions, including the bias, the root mean square error (RMSE) and the correlation between the true θ and the estimated θ. The recovery statistics were generally very similar for the true and estimated item banks. Poorest recovery was obtained for the $N(1.5, 0.2^2)$ θ distribution with three cutoffs.

Table 8. Marginal ATL and PCC Conditional on Number of Cutoffs, θ Distribution, and True and Estimated Item Banks

Number of Cutoffs and θ Distribution	ATL		PCC	
	True	Estimated	True	Estimated
(-0.44, 0.44)				
$N(0, 1^2)$	21.5	20.8	90.1	91.1
$N(-0.44, 0.2^2)$	28.6	28.4	76.6	78.5
$N(1, 0.2^2)$	15.7	14.6	98.2	98.2
(-1, 0, 1)				
$N(0, 1^2)$	26.4	25.2	87.6	86.2
$N(-0.5, 0.2^2)$	27.5	26.4	92.8	93.5
$N(1.5, 0.2^2)$	17.4	16.3	97.4	97.2

Table 8 displays the marginal ATL and PCC conditional on the θ distributions and true and estimated item banks. Using the estimated item parameters resulted in slightly shorter ATL (maximum 1.2 items) across all θ distributions. However, this led to a lower PCC in only two of the six θ distributions (maximum 1.4%), which means that the tests were not terminated prematurely.

Chapter 8. Discussion and Conclusions

Implications of Results

This study compared the classification accuracy and test length of various item selection algorithms and stopping rules that were designed to classify examinees into one of two or more categories. Consistent with the results of a previous study by Wang et al. (2021), mGLR was shown to be more effective in reducing test length than the long-established mSPRT, while sacrificing very little classification accuracy in a wide range of simulation conditions. The current study is the first to compare two new variants of stochastic curtailment methods (SC-MLE and SC-CI) to the originally proposed SC-Standard in multi-category CCT. All three SC-methods were able to reduce ATL without negatively affecting PCC. Under simulation, SC-MLE was shown to stop the tests more aggressively than SC-Standard and SC-CI, with only a slightly lower PCC. On average, SC-MLE required 3 to 4 fewer items than SC-Standard and SC-CI, but had only about 1% decrease in PCC. Overall, the results showed that the combination of mGLR and SC-MLE is the most ideal termination criterion with the best balance of test length and classification accuracy.

With respect to item selection methods, the current study demonstrated that FI-W was more efficient than FI-C and FI-E for all stopping rules. However, their differences were smaller than one item in most conditions, which suggests that the choice of item selection method has little practical importance. This conclusion is consistent with results of earlier studies that compared different item selection methods (e.g., Lau & Wang, 1999; Lin & Spray, 2000; Wouda & Eggen, 2009; van Groen et al., 2014). This might be due to the fact that the three item selection methods evaluated in this study are adaptive in

nature: the nearest cutoff depends on $\hat{\theta}$, and the likelihood function of $\hat{\theta}$ is based on the examinee's item responses. Since they assess the information provided by an item in a similar way, they tend to select the same item. Moreover, there was no evidence to support Thompson's (2009a) and Wang et al.'s (2021) hypothesis that FI-E is more suitable for stopping rules that use information regarding $\hat{\theta}$, such as mGLR, especially when the true θ is far away from the cutoffs. In fact, FI-E had the longest ATL regardless of the θ distribution.

In the CCT literature, there has also been speculation that if the item selection method will match item difficulty to the examinee's $\hat{\theta}$, then a wide range of difficulty parameters is necessary to cover a wide range of examinee ability (Thompson, 2007b). Unfortunately, detailed information regarding the item bank has typically not been reported in literature, except the mean and standard deviation of the item parameters. In light of this, the current study compared two item banks, one with a broad information function and one with a peaked information function. It was found that the ATL and PCC of CCT are sensitive to the item bank information function. Specifically, CCT using the item bank with a peaked information function was more efficient and accurate than the item bank with a broad information function, regardless of which item selection method was used. This result implies that the item bank information function for CCT does not necessarily have to include items with a very wide range of difficulty, if we are certain that the distribution of θ has a small standard deviation. In that case, the amount of information near the cutoff scores is more important.

The width of the indifference region (δ) plays an important role in the efficiency and accuracy of CCT. Previous studies typically specified δ to an arbitrary small

constant, or only examined a limited number of δs (e.g., Eggen, 1999; Finkelman, 2008; Finkelman, 2010; Huebner & Fina, 2015; Thompson, 2011). The current study extended previous studies by investigating four different levels of δ. The results suggested that δ had minor influence on PCC but considerable influence on ATL. In particular, increasing the value of δ resulted in a substantial decrease in ATL but only a small decrease in PCC. Therefore, δ should not be arbitrarily selected. Rather, a simulation such as the present study should be conducted to determine the value of δ that produces the shortest test lengths, while still maintaining the desired level of classification accuracy. A large δ also reduced the differences between stopping rules. A similar effect was also observed in Thompson (2011), who examined SPRT and GLR with two different δ levels in two-category CCT. This result is expected, as mGLR uses the same likelihood ratio as mSPRT to determine whether the test should be terminated, when $\hat{\theta}$ is inside an indifference region. When the indifference region is wide, there is a higher chance that $\hat{\theta}$ is inside an indifference region, and that mGLR and mSPRT will result in the same termination decision.

A somewhat interesting result was that when estimated item parameters were used, the ATL decreased slightly (a maximum of about one item) in some conditions, while the PCC had mixed results, increasing slightly in some conditions, and decreasing slightly in others. This corroborates the results in Patton et al. (2013), who found that classification accuracy improved for some θ levels when item calibration errors were introduced.

Limitations and Directions for Future Research

This study could be extended in a few directions. First, the current study, just as most CCT research, focused on the application of dichotomously scored items. Polytomous items can provide more information across a wider range of ability than dichotomously scored items (Birenbaum & Tatsuoka, 1987; Donoghue, 1994; Samejima, 1976), However, only a few studies have examined CCT with polytomous items (Gnambs & Batinic, 2011; Lau & Wang, 1998, 1999, 2000; Thompson, 2007), and they have used some of the older stopping rules such as ACI and SPRT. It will be interesting to evaluate the performance of recently developed stopping rules, such as GLR and SC, when polytomous items are used.

Second, future research should consider simulating item banks with different sets of parameter distributions, for example, to create a multimodal information function that is peaked at the cutoffs. Another possibility is to conduct the simulations with real item banks to gauge the generalizability of the current findings.

Third, the calibration sample size and the ratio of item bank size to test length have been shown to influence measurement quality (Hambleton et al., 1993; Hambleton & Jones, 1994; Olea et al., 2012; Patton et al., 2013; van der Linden & Glas, 2000). Future studies can vary these variables to study the effect of item calibration errors on multi-category CCT more comprehensively.

References

Armitage, P. (1950). Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society. Series B (Methodological), 12*(1), 137-144.

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

Bass, M., Morris, S., & Neapolitan, R. (2015). Utilizing multidimensional computer adaptive testing to mitigate burden with patient reported outcomes. In *AMIA Annual Symposium Proceedings* (Vol. 2015, p. 320). American Medical Informatics Association.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of general psychiatry, 4*(6), 561-571.

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

Birenbaum, M. & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats — it does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*(4), 385-395.

Camilli, G. (1994). Teacher's corner: origin of the scaling constant $d = 1.7$ in item response theory. *Journal of Educational and Behavioral Statistics, 19*(3), 293-295.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6).

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

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

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.). *Applications of item response theory* (pp. 175-195). Vancouver, BC: Educational Research Institute of British Columbia.

Cook, L. L., & Eignor, D. R. (1989). Using item response theory in test score equating. *International Journal of Educational Research, 13*(2), 161-173.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Donoghue, J. R. (1994). An empirical examination of the IRT information function of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement, 31*, 295-311.

Eggen, T. J. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*(3), 249-261.

Eggen, T. J. (2009). Three-category adaptive classification testing. In W. J. van der Linden & C. A.W. Glas (Eds.), *Elements of adaptive testing* (pp. 373–387). Springer.

Eggen, T. & Straetmans, G. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*(5), 713–734.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Ferguson, R. L. (1969). *Computer-assisted criterion-referenced measurement* (Working Paper No. 41). Pittsburgh: University of Pittsburgh, Learning and Research Development Center. (Eric Document Reproduction Series No. ED 037 089).

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics, 33*(4), 442-463.

Finkelman, M. D. (2010). Variations on stochastic curtailment in sequential mastery testing. *Applied Psychological Measurement, 34*(1), 27-45.

Finkelman, M. D., He, Y., Kim, W., & Lai, A. M. (2011). Stochastic curtailment of health questionnaires: A method to reduce respondent burden. *Statistics in Medicine, 30,* 1989-2004.

Frick, T. W. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research, 8*(2), 187-213.

Ghosh, B. K. (1970). *Sequential tests of statistical hypotheses*. Reading, MA: Addison-Wesley.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric services, 59*(4), 361-368.

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*(6), 100.06-1022.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

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

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement, 30*(2), 143-155.

Hambleton, R. K., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education, 7*(3), 171-186.

Hu, B. H., & Shih, C. L. (2021). Revising and evaluating the expected likelihood ratio as a new item selection strategy in unidimensional computerized classification tests. *International Journal of Intelligent Technologies & Applied Statistics, 14*(2).

Huebner, A. R., & Fina, A. D. (2015). The stochastically curtailed generalized likelihood ratio: A new termination criterion for variable-length computerized classification tests. *Behavior Research Methods, 47*(2), 549-561.

Huebner, A. R., & Finkelman, M. D. (2016). On computing the key probability in the stochastically curtailed sequential probability ratio test. *Applied Psychological Measurement, 40*(2), 142-156.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*(3), 249-260.

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 test theory and computerized adaptive testing* (pp. 257-283). New York: Academic Press.

Lan, K. K. G., Simon, R., & Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials. *Sequential Analysis, 1*(3), 207-219.

Lau, C. A., & Wang, T. (1998). *Comparing and combining dichotomous and polytomous items with SPRT procedure in computerized classification testing*. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Lau, C. A., & Wang, T. (1999, April). *Computerized classification testing under practical constraints with a polytomous model*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Lin, C. J., & Spray, J. (2000). *Effects of item-selection criteria on classification testing with the sequential probability ratio test*. ACT Research Report Series.

Lin, C. J. (2011). Item selection criteria with practical constraints for computerized classification testing. *Educational and Psychological Measurement, 71*(1), 20-36.

Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph No.1). Iowa City. IA: Psychometric Society.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three- parameter logistic model. *Educational and Psychological Measurement, 28*(4), 989-1020.

Lord. F M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale. NJ: Lawrence Erlbaum.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series, 1992*(1), i-30.

Nydick, S. W. (2013). *Multidimensional mastery testing with CAT* [Unpublished doctoral dissertation]. University of Minnesota.

Olea, J., Barrada, J. R., Abad, F. J., Ponsoda, V., & Cuevas, L. (2012). Computerized adaptive testing: The capitalization on chance problem. *The Spanish Journal of Psychology, 15*(1), 424-441.

Patton, J. M., Cheng, Y., Yuan, K. H., & Diao, Q. (2013). The influence of item calibration error on variable-length computerized adaptive testing. *Applied Psychological Measurement, 37*(1), 24-40.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test*. Copenhagen: The Danish Institute for Educational Research.

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.

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer, New York, NY.

Ren, H., Huang, G, Y., & Chen, P. (2022). Types, characteristics and application of termination rules in computerized classification testing. *Advances in Psychological Science, 30*(5), 1168.

Rudner, L. M. (2002, April). *An examination of decision-theory adaptive testing procedures* [Paper presentation]. American Educational Research Association, New Orleans, LA.

Sahin, A., & Anil, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice, 17*(1), 321-335.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society.

Samejima, F. (1976). Graded response model of the latent trait theory and tailored testing. In *C. K. Clark (Ed.), Proceedings of the first Conference on Computerized Adaptive Testing (pp. 5-17). Washington, DC: U.S. Government Printing Office*.

Shannon, C.E., & Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press, Urbana, Illinois.

Sie, H., Finkelman, M. D., Bartroff, J., & Thompson, N. A. (2015). Stochastic curtailment in adaptive mastery testing: Improving the efficiency of confidence interval-based stopping rules. *Applied Psychological Measurement, 39*(4), 278-292.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237-247.

Sobel, M. & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *The Annals of Mathematical Statistics, 20*(4), 502–522.

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

Spray, J. A., & Reckase, M. D. (1994, April). *The selection of test items for decision making with a computer adaptive test* [Paper presentation]. American Educational Research Association, 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*(4), 405-414.

Stroud, A.H., & Sechrest, D. (1966). *Gaussian quadrature formulas*. Englewood Cliffs: Prentice-Hall.

Thompson, N. A. (2007a). *A comparison of two methods of polytomous computerized classification testing for multiple cutscores* [Unpublished doctoral dissertation]. University of Minnesota.

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

Thompson, N. A. (2009a). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*(5), 778-793.

Thompson, N. A. (2009b). Utilizing the generalized likelihood ratio as a termination criterion. In *GMAC conference on computerized adaptive testing*, Minneapolis, MN.

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research, and Evaluation, 16*(1), 4.

van der Linden, W. J., & Glas, C. A. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education, 13*(1), 35-53.

van Groen, M. M., Eggen, T. J., & Veldkamp, B. P. (2014). Item selection methods based on multiple objective approaches for classifying respondents into multiple levels. *Applied Psychological Measurement, 38*(3), 187-200.

Veerkamp, W. J., & Berger, M. P. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*(2), 203-226.

Vos, H. J. (1998). Optimal sequential rules for computer-based instruction. *Journal of Educational Computing Research, 19*, 133-154.

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

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22-29.

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

Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. Psychometrika, 80(2), 428-449.

Wang, C., Chen, P., & Huebner, A. (2021). Stopping rules for multi-category computerized classification testing. *British Journal of Mathematical and Statistical Psychology, 74*(2), 184-202.

Wang, Z., Wang, C., & Weiss, D. J. (2022). Termination criteria for grid multiclassification adaptive testing with multidimensional polytomous items. *Applied Psychological Measurement, 46*(7), 551-570.

Weiss, D. J. (Ed.). (1983). *New horizons in testing*. New York: Academic Pre5s.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361-375.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences, 2*(1), 1-27.

Weissman, A. (2004). *Mutual information item selection in multiple-category classification CAT* [Paper presentation]. National Council for Measurement in Education, San Diego, CA.

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement, 67*(1), 41-58.

Wouda, J. T., & Eggen, T. J. H. M. (2009). Computerized classification testing in more than two categories by using stochastic curtailment. In *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*.

Appendix A

Table A-1. Average Test Length for Cutoffs = (-0.44, 0.44) and Broad Item Bank with True Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	41.0	43.2	40.8	27.9	29.7	27.7	20.5	22.3	20.9	16.2	18.0	16.4
mGLR	30.9	30.9	30.8	25.6	26.0	25.6	20.3	21.5	20.7	16.2	17.8	16.4
mSPRT + SC-Standard	31.8	31.2	31.7	27.2	28.1	27.0	20.4	22.0	20.8	16.1	17.9	16.3
mSPRT + SC-MLE	24.0	23.7	23.8	20.6	20.5	20.4	17.0	17.6	17.3	14.9	16.0	14.8
mSPRT + SC-CI	31.6	29.6	31.4	27.1	26.2	26.9	20.4	21.6	20.8	16.1	17.8	16.3
mGLR + SC-Standard	28.6	28.3	28.6	25.0	25.1	24.9	20.2	21.4	20.6	16.1	17.8	16.3
mGLR + SC-MLE	22.5	22.2	22.3	19.9	19.7	19.7	16.9	17.6	17.2	14.9	16.0	14.8
mGLR + SC-CI	28.8	28.1	28.6	24.9	25.0	24.9	20.1	21.3	20.6	16.1	17.8	16.3
$N(-0.44, 0.2^2)$												
mSPRT	49.3	49.4	49.3	39.7	40.4	39.5	28.8	30.5	29.0	21.9	23.9	22.0
mGLR	43.5	43.7	43.4	37.4	37.5	37.1	28.6	30.0	28.8	22.0	24.0	22.0
mSPRT + SC-Standard	41.7	41.2	41.7	37.7	38.2	37.5	28.5	30.1	28.7	21.8	23.8	21.9
mSPRT + SC-MLE	31.5	30.6	31.2	27.6	27.1	27.1	22.6	23.1	22.5	19.4	20.4	19.1
mSPRT + SC-CI	41.7	40.7	41.8	37.5	37.0	37.4	28.5	30.1	28.7	21.8	23.8	21.9
mGLR + SC-Standard	39.5	39.5	39.4	35.7	35.8	35.4	28.3	29.7	28.5	21.9	23.9	21.9
mGLR + SC-MLE	29.9	29.2	29.4	26.8	26.3	26.1	22.5	23.1	22.5	19.4	20.4	19.1
mGLR + SC-CI	39.8	39.3	39.7	35.6	35.6	35.3	28.2	29.7	28.4	21.9	23.9	21.9
$N(1, 0.2^2)$												
mSPRT	37.4	42.1	37.0	20.3	23.8	20.2	15.1	17.2	15.0	13.0	14.2	13.0
mGLR	21.8	22.2	21.2	17.7	18.8	17.5	14.9	16.0	14.8	13.0	14.0	13.1
mSPRT + SC-Standard	22.2	22.4	21.8	19.6	22.5	19.5	15.0	17.0	14.9	12.9	14.2	13.0
mSPRT + SC-MLE	12.7	12.5	12.5	12.7	12.5	12.5	12.4	12.4	12.2	12.1	12.1	11.9
mSPRT + SC-CI	20.2	18.8	19.9	19.4	18.9	19.2	14.9	16.2	14.9	12.9	14.0	13.0
mGLR + SC-Standard	19.1	19.3	18.6	17.1	18.1	17.0	14.8	15.9	14.7	12.9	13.9	13.0
mGLR + SC-MLE	12.7	12.5	12.5	12.7	12.5	12.4	12.4	12.4	12.2	12.1	12.1	11.9
mGLR + SC-CI	18.9	18.8	18.4	17.1	18.0	16.9	14.8	15.9	14.7	12.9	13.9	13.0

Table A-1. Percentage of Correct Classification for Cutoffs = (-0.44, 0.44) and Broad Item Bank with True Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	90.2	90.3	90.6	90.1	90.3	90.5	90.0	89.2	89.7	89.3	88.0	89.4
mGLR	90.1	89.8	90.5	89.9	89.5	90.3	89.9	88.9	89.7	89.3	88.0	89.4
mSPRT + SC-Standard	90.1	89.9	90.5	90.1	90.1	90.5	90.0	89.2	89.7	89.3	88.0	89.4
mSPRT + SC-MLE	87.8	86.4	88.0	87.8	86.6	88.0	88.2	86.9	87.7	88.2	86.7	88.0
mSPRT + SC-CI	90.1	89.4	90.5	90.1	89.6	90.5	90.0	89.1	89.7	89.3	88.0	89.4
mGLR + SC-Standard	90.0	89.4	90.4	89.9	89.4	90.3	89.9	88.9	89.7	89.3	88.0	89.4
mGLR + SC-MLE	87.8	86.4	88.0	87.9	86.6	88.1	88.1	86.8	87.7	88.2	86.7	88.0
mGLR + SC-CI	90.1	89.4	90.5	89.9	89.4	90.3	89.9	88.9	89.7	89.3	88.0	89.4
$N(-0.44, 0.2^2)$												
mSPRT	77.0	77.0	77.0	77.1	77.1	77.1	76.2	75.9	76.5	74.3	74.5	74.9
mGLR	77.1	76.7	77.1	76.9	76.4	76.8	76.3	75.8	76.6	74.5	74.6	74.9
mSPRT + SC-Standard	77.0	76.8	77.0	77.1	77.1	77.1	76.3	75.9	76.6	74.3	74.6	74.9
mSPRT + SC-MLE	73.8	73.6	73.9	73.8	73.8	73.9	73.2	73.2	73.8	72.5	72.5	73.3
mSPRT + SC-CI	76.9	76.6	76.9	77.0	76.8	77.0	76.3	75.7	76.6	74.3	74.5	74.9
mGLR + SC-Standard	77.1	76.6	77.1	76.9	76.4	76.8	76.4	75.8	76.7	74.5	74.7	74.9
mGLR + SC-MLE	73.8	73.6	73.9	73.8	73.5	73.8	73.2	73.2	73.8	72.5	72.5	73.3
mGLR + SC-CI	77.0	76.6	77.0	76.8	76.5	76.7	76.4	75.8	76.7	74.5	74.7	74.9
$N(1, 0.2^2)$												
mSPRT	98.5	98.5	98.5	98.5	98.5	98.5	97.8	98.1	97.8	97.0	97.4	97.2
mGLR	98.5	98.6	98.5	98.5	98.4	98.3	97.9	98.1	97.9	97.0	97.4	97.3
mSPRT + SC-Standard	98.5	98.6	98.5	98.5	98.5	98.5	97.8	98.1	97.8	97.0	97.4	97.2
mSPRT + SC-MLE	98.7	99.1	98.9	98.7	99.0	98.9	97.9	98.6	98.1	97.1	97.8	97.3
mSPRT + SC-CI	98.5	98.6	98.5	98.5	98.5	98.5	97.8	98.1	97.8	97.0	97.4	97.2
mGLR + SC-Standard	98.5	98.6	98.5	98.5	98.4	98.3	97.9	98.1	97.9	97.0	97.4	97.3
mGLR + SC-MLE	98.7	99.1	98.9	98.7	98.9	98.7	98.0	98.6	98.2	97.1	97.8	97.4
mGLR + SC-CI	98.5	98.6	98.5	98.5	98.4	98.3	97.9	98.1	97.9	97.0	97.4	97.3

Table A-2. Average Test Length for Cutoffs = (-0.44, 0.44) and Peaked Item Bank with True Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	37.7	39.0	37.4	23.8	25.5	23.3	17.4	18.7	17.3	13.8	15.3	13.9
mGLR	28.0	28.8	27.7	22.5	23.5	22.1	17.5	18.5	17.3	13.8	15.3	13.9
mSPRT + SC-Standard	31.6	31.4	31.2	23.4	25.0	22.9	17.4	18.6	17.2	13.8	15.3	13.9
mSPRT + SC-MLE	23.5	22.9	23.7	18.4	18.3	18.6	15.3	15.6	15.4	13.1	14.0	13.4
mSPRT + SC-CI	30.9	29.4	30.4	23.4	24.4	22.9	17.4	18.6	17.2	13.8	15.3	13.9
mGLR + SC-Standard	26.3	26.8	26.0	22.1	23.0	21.7	17.5	18.4	17.2	13.8	15.3	13.9
mGLR + SC-MLE	20.7	20.4	21.0	18.0	17.8	18.1	15.3	15.6	15.4	13.1	13.9	13.4
mGLR + SC-CI	26.3	26.8	26.1	22.1	23.0	21.7	17.5	18.4	17.2	13.8	15.3	13.9
$N(-0.44, 0.2^2)$												
mSPRT	47.2	47.8	46.9	33.0	34.4	32.7	22.5	23.7	22.0	16.6	18.1	16.7
mGLR	39.6	40.4	39.3	31.4	32.4	31.5	22.5	23.6	21.9	16.6	18.1	16.7
mSPRT + SC-Standard	40.8	40.7	40.6	32.2	33.5	31.9	22.4	23.6	21.9	16.6	18.1	16.7
mSPRT + SC-MLE	30.1	29.9	30.0	23.6	24.3	23.5	18.7	19.6	18.3	15.5	16.6	15.4
mSPRT + SC-CI	40.6	39.9	40.3	32.2	33.2	31.9	22.4	23.6	21.9	16.6	18.1	16.7
mGLR + SC-Standard	36.6	36.8	36.3	30.7	31.5	30.7	22.3	23.5	21.8	16.5	18.1	16.7
mGLR + SC-MLE	27.0	27.1	26.9	22.9	23.7	22.9	18.7	19.6	18.3	15.5	16.6	15.4
mGLR + SC-CI	36.8	36.9	36.4	30.6	31.5	30.7	22.3	23.5	21.8	16.5	18.1	16.7
$N(1, 0.2^2)$												
mSPRT	32.2	34.8	31.9	17.6	19.3	16.9	13.7	14.5	13.4	11.9	12.6	11.9
mGLR	19.4	19.8	19.4	16.3	16.5	15.8	13.7	14.0	13.4	11.9	12.6	11.9
mSPRT + SC-Standard	22.6	22.5	22.2	17.4	19.1	16.7	13.6	14.4	13.3	11.9	12.6	11.9
mSPRT + SC-MLE	12.0	11.9	11.9	11.9	11.9	11.8	11.7	11.6	11.6	11.3	11.4	11.3
mSPRT + SC-CI	20.0	17.9	19.3	17.4	17.6	16.7	13.6	14.3	13.3	11.9	12.6	11.9
mGLR + SC-Standard	17.8	18.3	17.8	16.1	16.3	15.6	13.6	14.0	13.4	11.9	12.6	11.9
mGLR + SC-MLE	11.9	11.9	11.9	11.9	11.8	11.8	11.7	11.7	11.6	11.3	11.4	11.3
mGLR + SC-CI	17.7	17.8	17.7	16.0	16.3	15.5	13.6	14.0	13.3	11.9	12.6	11.9

Table A-3. Percentage of Correct Classification for Cutoffs = (-0.44, 0.44) and Peaked Item Bank with True Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	92.2	92.2	92.2	91.9	92.3	91.9	91.5	91.1	91.6	89.2	90.0	89.9
mGLR	92.0	92.2	92.2	91.8	91.9	92.0	91.5	90.7	91.6	89.2	90.0	89.9
mSPRT + SC-Standard	92.3	92.3	92.2	92.0	92.3	91.9	91.5	91.1	91.6	89.2	90.0	89.9
mSPRT + SC-MLE	91.0	89.7	91.8	90.6	89.5	91.4	90.2	88.7	90.8	88.6	88.5	89.6
mSPRT + SC-CI	92.3	92.3	92.2	92.0	92.1	91.9	91.5	91.0	91.6	89.2	90.0	89.9
mGLR + SC-Standard	92.1	92.2	92.2	91.9	92.0	92.0	91.5	90.7	91.6	89.2	90.0	89.9
mGLR + SC-MLE	90.9	89.7	91.8	90.6	89.6	91.4	90.3	88.7	90.8	88.6	88.5	89.6
mGLR + SC-CI	92.1	92.3	92.2	91.9	92.0	92.0	91.5	90.7	91.6	89.2	90.0	89.9
$N(-0.44, 0.2^2)$												
mSPRT	79.7	78.5	80.3	79.7	78.7	80.1	79.1	77.3	78.2	76.5	75.6	76.6
mGLR	79.8	78.1	80.5	79.1	78.0	79.9	79.1	77.3	78.1	76.4	75.6	76.6
mSPRT + SC-Standard	79.8	78.7	80.1	79.7	78.8	80.0	79.1	77.3	78.2	76.5	75.6	76.6
mSPRT + SC-MLE	76.5	75.8	77.3	76.3	75.7	76.9	76.6	74.7	76.0	75.7	73.8	75.3
mSPRT + SC-CI	79.7	78.2	80.2	79.7	78.4	80.0	79.1	77.3	78.2	76.5	75.6	76.6
mGLR + SC-Standard	79.9	78.4	80.3	79.1	78.1	79.8	79.1	77.3	78.1	76.4	75.6	76.6
mGLR + SC-MLE	76.6	75.8	77.4	76.0	75.4	76.7	76.6	74.7	75.9	75.6	73.8	75.3
mGLR + SC-CI	79.8	78.3	80.4	79.1	78.1	79.8	79.1	77.3	78.1	76.4	75.6	76.6
$N(1, 0.2^2)$												
mSPRT	98.8	98.7	98.7	98.8	98.7	98.6	98.2	98.4	98.0	97.3	97.4	97.5
mGLR	98.8	98.7	98.7	98.7	98.8	98.5	98.2	98.4	98.0	97.4	97.5	97.5
mSPRT + SC-Standard	98.8	98.7	98.7	98.8	98.7	98.6	98.2	98.4	98.0	97.3	97.4	97.5
mSPRT + SC-MLE	99.1	99.1	99.0	99.1	99.1	98.9	98.5	98.7	98.3	97.6	97.7	97.9
mSPRT + SC-CI	98.8	98.7	98.7	98.8	98.7	98.6	98.2	98.4	98.0	97.3	97.4	97.5
mGLR + SC-Standard	98.8	98.7	98.7	98.7	98.8	98.5	98.2	98.4	98.0	97.4	97.5	97.5
mGLR + SC-MLE	99.1	99.1	99.0	99.0	99.1	98.8	98.5	98.7	98.3	97.7	97.8	97.9
mGLR + SC-CI	98.8	98.7	98.7	98.7	98.8	98.5	98.2	98.4	98.0	97.4	97.5	97.5

Table A-4. Average Test Length for Cutoffs = (-0.44, 0.44) and Broad Item Bank with Estimated Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	39.6	41.2	39.4	26.2	27.2	25.9	19.3	20.2	19.1	15.3	16.5	15.3
mGLR	30.4	30.0	30.1	24.6	24.7	24.5	19.2	19.9	19.1	15.2	16.3	15.3
mSPRT + SC-Standard	31.8	31.2	31.5	25.6	26.3	25.3	19.2	20.1	19.0	15.3	16.4	15.2
mSPRT + SC-MLE	24.4	24.1	24.5	20.1	20.0	20.0	16.7	17.0	16.6	14.6	15.2	14.5
mSPRT + SC-CI	31.6	29.9	31.3	25.6	25.5	25.2	19.2	20.0	19.0	15.3	16.4	15.2
mGLR + SC-Standard	28.1	27.8	27.8	24.0	24.1	23.9	19.1	19.8	19.0	15.2	16.3	15.3
mGLR + SC-MLE	22.3	21.9	22.3	19.6	19.5	19.5	16.5	17.0	16.6	14.4	15.1	14.5
mGLR + SC-CI	28.1	27.6	27.8	24.0	24.1	23.8	19.1	19.8	19.0	15.2	16.3	15.3
$N(-0.44, 0.2^2)$												
mSPRT	49.0	49.0	49.0	37.3	38.3	36.9	27.3	28.9	27.4	20.4	22.7	20.6
mGLR	42.1	42.3	41.9	35.4	35.7	35.2	27.1	28.8	27.4	20.2	22.6	20.6
mSPRT + SC-Standard	41.4	40.8	41.3	36.0	36.6	35.6	27.1	28.7	27.2	20.4	22.7	20.6
mSPRT + SC-MLE	32.0	32.3	32.0	27.3	27.9	27.1	22.5	23.9	22.6	19.0	21.0	19.2
mSPRT + SC-CI	41.6	40.6	41.6	35.9	36.0	35.6	27.1	28.6	27.2	20.3	22.7	20.6
mGLR + SC-Standard	38.4	38.1	38.1	34.2	34.2	34.0	26.9	28.5	27.1	20.1	22.6	20.5
mGLR + SC-MLE	29.7	30.1	29.6	26.4	27.0	26.5	22.3	23.9	22.6	18.8	20.9	19.2
mGLR + SC-CI	38.7	38.3	38.4	34.2	34.1	34.0	26.9	28.5	27.1	20.1	22.6	20.5
$N(1, 0.2^2)$												
mSPRT	32.8	36.4	32.2	17.1	19.3	16.8	13.5	14.2	13.2	12.0	12.2	11.8
mGLR	19.6	19.4	19.3	15.5	16.1	15.4	13.5	13.7	13.3	11.9	12.1	11.8
mSPRT + SC-Standard	20.4	20.5	19.7	16.7	18.8	16.5	13.4	14.1	13.2	12.0	12.2	11.8
mSPRT + SC-MLE	11.8	11.7	11.8	11.7	11.7	11.8	11.6	11.4	11.6	11.4	11.2	11.4
mSPRT + SC-CI	18.0	17.0	17.5	16.7	17.0	16.4	13.4	13.9	13.2	12.0	12.2	11.8
mGLR + SC-Standard	17.4	17.4	17.0	15.2	15.8	15.1	13.4	13.6	13.2	11.9	12.1	11.8
mGLR + SC-MLE	11.8	11.7	11.8	11.7	11.7	11.8	11.6	11.4	11.6	11.4	11.2	11.4
mGLR + SC-CI	17.2	17.0	16.7	15.2	15.7	15.1	13.4	13.6	13.2	11.9	12.1	11.8

Table A-5. Percentage of Correct Classification for Cutoffs = (-0.44, 0.44) and Broad Item Bank with Estimated Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	91.6	92.1	91.5	91.5	92.0	91.4	90.9	91.5	91.3	89.4	90.2	89.8
mGLR	91.4	92.0	91.4	91.1	91.8	91.2	90.8	91.5	91.3	89.2	90.0	89.8
mSPRT + SC-Standard	91.6	92.0	91.4	91.5	91.9	91.4	90.9	91.4	91.3	89.4	90.1	89.8
mSPRT + SC-MLE	89.7	89.1	89.6	89.7	89.2	89.6	89.9	89.1	89.7	89.6	88.7	89.2
mSPRT + SC-CI	91.6	91.8	91.5	91.5	91.8	91.4	90.9	91.4	91.3	89.4	90.1	89.8
mGLR + SC-Standard	91.4	91.9	91.3	91.1	91.7	91.2	90.8	91.4	91.3	89.2	89.9	89.8
mGLR + SC-MLE	89.7	89.1	89.6	89.8	89.1	89.6	89.8	89.1	89.7	89.4	88.5	89.2
mGLR + SC-CI	91.5	91.8	91.5	91.1	91.7	91.2	90.8	91.4	91.3	89.2	89.9	89.8
$N(-0.44, 0.2^2)$												
mSPRT	78.7	79.3	78.5	78.8	78.8	78.6	77.6	77.5	77.5	76.4	76.2	76.1
mGLR	78.7	79.1	78.4	78.4	78.5	78.3	77.3	77.4	77.5	76.3	76.3	76.1
mSPRT + SC-Standard	78.5	79.2	78.3	78.7	78.9	78.5	77.5	77.6	77.4	76.2	76.2	76.0
mSPRT + SC-MLE	77.3	76.6	77.3	77.3	76.4	77.3	77.0	76.3	76.8	75.6	75.2	75.5
mSPRT + SC-CI	78.7	79.0	78.5	78.7	78.8	78.5	77.5	77.6	77.4	76.1	76.2	76.0
mGLR + SC-Standard	78.5	79.0	78.2	78.3	78.6	78.2	77.2	77.5	77.4	76.1	76.3	76.0
mGLR + SC-MLE	77.3	76.6	77.3	77.3	76.4	77.3	76.7	76.3	76.7	75.5	75.3	75.5
mGLR + SC-CI	78.7	79.0	78.4	78.3	78.6	78.2	77.2	77.5	77.4	76.0	76.3	76.0
$N(1, 0.2^2)$												
mSPRT	98.6	98.1	98.6	98.6	98.1	98.6	98.5	97.9	98.4	98.1	97.4	98.0
mGLR	98.6	98.1	98.6	98.7	98.1	98.6	98.5	97.8	98.4	98.0	97.4	98.0
mSPRT + SC-Standard	98.7	98.3	98.7	98.7	98.2	98.7	98.5	97.9	98.4	98.1	97.4	98.0
mSPRT + SC-MLE	98.9	98.6	98.9	98.8	98.6	98.9	98.6	98.3	98.5	98.2	97.6	98.1
mSPRT + SC-CI	98.7	98.2	98.7	98.7	98.2	98.7	98.5	97.9	98.4	98.1	97.4	98.0
mGLR + SC-Standard	98.7	98.3	98.7	98.7	98.1	98.7	98.5	97.8	98.4	98.0	97.4	98.0
mGLR + SC-MLE	98.9	98.6	98.9	98.8	98.5	98.9	98.6	98.2	98.5	98.1	97.6	98.1
mGLR + SC-CI	98.7	98.2	98.7	98.7	98.1	98.7	98.5	97.8	98.4	98.0	97.4	98.0

Table A-6. Average Test Length for Cutoffs = (-0.44, 0.44) and Peaked Item Bank with Estimated Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	36.7	37.8	36.3	23.0	24.1	22.4	16.8	18.0	16.3	14.1	14.9	13.8
mGLR	27.7	27.8	27.3	21.9	22.5	21.5	16.6	17.8	16.3	14.1	14.8	13.8
mSPRT + SC-Standard	31.4	30.6	30.9	22.6	23.8	22.1	16.8	17.9	16.3	14.1	14.9	13.8
mSPRT + SC-MLE	22.9	22.7	23.1	17.7	18.0	17.8	14.6	15.2	14.7	13.2	13.7	13.3
mSPRT + SC-CI	30.5	28.6	30.0	22.6	23.2	22.1	16.7	17.9	16.3	14.1	14.9	13.8
mGLR + SC-Standard	26.1	26.1	25.8	21.6	22.2	21.2	16.6	17.8	16.3	14.1	14.8	13.8
mGLR + SC-MLE	20.2	20.2	20.5	17.3	17.5	17.3	14.5	15.2	14.6	13.1	13.6	13.3
mGLR + SC-CI	26.1	25.9	25.8	21.6	22.2	21.2	16.6	17.8	16.3	14.1	14.8	13.8
$N(-0.44, 0.2^2)$												
mSPRT	47.3	47.8	47.1	33.6	34.1	32.8	23.3	24.0	22.8	17.6	18.5	17.2
mGLR	40.3	40.5	40.1	32.1	32.4	31.5	23.2	23.9	22.8	17.6	18.4	17.2
mSPRT + SC-Standard	41.2	41.2	41.0	32.8	33.4	32.1	23.2	23.9	22.7	17.6	18.5	17.2
mSPRT + SC-MLE	31.2	29.9	30.9	24.7	23.9	24.2	20.1	19.1	19.7	16.2	16.6	16.4
mSPRT + SC-CI	41.0	40.4	40.7	32.8	33.2	32.1	23.2	23.9	22.7	17.6	18.5	17.2
mGLR + SC-Standard	37.1	37.1	36.9	31.4	31.7	30.8	23.1	23.8	22.7	17.6	18.4	17.2
mGLR + SC-MLE	28.0	26.9	27.9	24.1	23.0	23.5	19.9	19.1	19.7	16.2	16.5	16.4
mGLR + SC-CI	37.3	37.2	37.0	31.3	31.6	30.8	23.1	23.8	22.7	17.6	18.4	17.2
$N(1, 0.2^2)$												
mSPRT	30.2	32.1	29.5	16.3	17.9	15.9	13.0	13.5	12.8	11.7	12.1	11.7
mGLR	19.2	19.0	19.0	15.6	15.9	15.4	13.0	13.4	12.8	11.7	12.0	11.7
mSPRT + SC-Standard	22.3	21.5	21.4	16.1	17.7	15.8	13.0	13.5	12.8	11.7	12.1	11.7
mSPRT + SC-MLE	12.0	11.8	11.8	11.8	11.8	11.7	11.6	11.6	11.5	11.2	11.3	11.2
mSPRT + SC-CI	19.7	17.2	18.9	16.1	16.7	15.7	13.0	13.5	12.8	11.7	12.1	11.7
mGLR + SC-Standard	17.6	17.5	17.4	15.4	15.7	15.2	12.9	13.3	12.8	11.7	12.0	11.7
mGLR + SC-MLE	12.0	11.8	11.8	11.8	11.8	11.7	11.6	11.6	11.5	11.2	11.2	11.2
mGLR + SC-CI	17.5	17.2	17.3	15.4	15.7	15.2	12.9	13.3	12.8	11.7	12.0	11.7

Table A-7. Percentage of Correct Classification for Cutoffs = (-0.44, 0.44) and Peaked Item Bank with Estimated Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	93.2	93.7	93.4	93.3	93.0	93.2	91.6	92.5	91.8	90.2	91.4	90.5
mGLR	93.2	93.6	93.3	93.1	93.1	92.9	91.6	92.3	91.7	90.3	91.4	90.5
mSPRT + SC-Standard	93.2	93.4	93.3	93.2	92.8	93.1	91.6	92.3	91.8	90.2	91.4	90.5
mSPRT + SC-MLE	90.3	89.7	91.1	90.3	89.2	90.9	89.6	89.2	90.3	88.8	88.9	90.1
mSPRT + SC-CI	93.2	93.3	93.3	93.2	92.8	93.1	91.6	92.3	91.8	90.2	91.4	90.5
mGLR + SC-Standard	93.2	93.3	93.2	93.0	92.9	92.8	91.6	92.1	91.7	90.3	91.4	90.5
mGLR + SC-MLE	90.3	89.6	91.1	90.3	89.3	90.6	89.6	89.1	90.3	88.9	89.0	90.1
mGLR + SC-CI	93.2	93.2	93.2	93.0	92.9	92.8	91.6	92.1	91.7	90.3	91.4	90.5
$N(-0.44, 0.2^2)$												
mSPRT	80.2	79.7	79.5	80.7	79.9	80.0	80.3	79.5	80.7	78.8	78.7	78.9
mGLR	80.2	80.0	79.5	80.8	79.6	79.7	80.4	79.5	80.7	78.8	78.6	78.9
mSPRT + SC-Standard	80.4	80.2	80.1	80.7	80.1	80.3	80.3	79.6	80.8	78.8	78.6	78.9
mSPRT + SC-MLE	79.2	77.8	80.1	79.3	77.9	80.2	78.7	77.7	79.8	77.6	76.9	78.7
mSPRT + SC-CI	80.2	80.2	79.8	80.7	80.1	80.1	80.3	79.5	80.8	78.8	78.6	78.8
mGLR + SC-Standard	80.4	80.5	80.1	80.8	79.8	80.0	80.4	79.6	80.8	78.8	78.5	78.9
mGLR + SC-MLE	79.2	77.9	80.1	79.2	77.9	80.1	78.8	77.8	79.8	77.6	76.8	78.7
mGLR + SC-CI	80.2	80.3	79.9	80.8	79.8	79.8	80.4	79.5	80.8	78.8	78.5	78.8
$N(1, 0.2^2)$												
mSPRT	98.7	98.4	98.7	98.7	98.4	98.8	98.2	97.5	97.9	96.9	96.4	96.7
mGLR	98.8	98.4	98.8	98.7	98.4	98.8	98.2	97.6	97.9	96.9	96.3	96.7
mSPRT + SC-Standard	98.8	98.4	98.9	98.7	98.4	98.8	98.2	97.5	97.9	96.9	96.4	96.7
mSPRT + SC-MLE	99.1	98.7	99.0	98.9	98.7	98.9	98.3	97.8	98.0	97.0	96.6	96.7
mSPRT + SC-CI	98.8	98.4	98.8	98.7	98.4	98.8	98.2	97.5	97.9	96.9	96.4	96.7
mGLR + SC-Standard	98.8	98.4	98.9	98.7	98.4	98.8	98.2	97.6	97.9	96.9	96.3	96.7
mGLR + SC-MLE	99.1	98.7	99.0	98.8	98.6	98.9	98.3	97.8	98.0	97.0	96.5	96.7
mGLR + SC-CI	98.8	98.4	98.8	98.7	98.4	98.8	98.2	97.6	97.9	96.9	96.3	96.7

Table A-8. Average Test Length for Cutoffs = (-1, 0, 1) and Broad Item Bank with True Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	47.0	47.7	46.9	33.5	34.8	32.9	23.6	25.3	23.3	18.2	19.5	18.1
mGLR	37.8	38.2	37.6	30.4	31.7	30.3	23.1	24.6	23.2	17.9	19.4	18.1
mSPRT + SC-Standard	42.0	41.9	41.9	32.7	33.8	32.1	23.5	25.2	23.2	18.1	19.5	18.0
mSPRT + SC-MLE	37.3	36.0	37.0	28.6	28.0	27.9	21.6	21.7	21.1	17.3	17.9	17.2
mSPRT + SC-CI	42.0	41.1	42.0	32.7	32.9	32.1	23.5	25.1	23.2	18.1	19.5	18.0
mGLR + SC-Standard	36.0	36.3	35.8	29.8	30.9	29.6	23.0	24.5	23.1	17.9	19.4	18.0
mGLR + SC-MLE	32.1	31.1	31.7	26.5	26.4	26.1	21.1	21.4	21.0	17.1	17.8	17.1
mGLR + SC-CI	36.0	36.1	35.8	29.8	30.9	29.6	23.0	24.5	23.1	17.9	19.4	18.0
$N(-0.5, 0.2^2)$												
mSPRT	50.0	50.0	50.0	35.1	35.2	34.4	25.0	25.2	23.8	19.2	19.2	18.4
mGLR	40.4	40.8	40.2	31.7	32.7	31.5	24.0	24.5	23.4	18.9	19.2	18.4
mSPRT + SC-Standard	49.6	49.5	49.6	35.0	35.0	34.2	25.0	25.1	23.7	19.2	19.2	18.4
mSPRT + SC-MLE	47.9	47.3	48.0	33.5	33.1	32.6	23.9	23.8	22.7	18.6	18.6	17.8
mSPRT + SC-CI	49.6	49.4	49.6	35.0	34.9	34.2	25.0	25.1	23.7	19.2	19.2	18.4
mGLR + SC-Standard	40.2	40.5	40.0	31.6	32.5	31.4	24.0	24.5	23.3	18.9	19.2	18.4
mGLR + SC-MLE	38.6	38.5	38.4	30.2	30.8	30.0	23.0	23.2	22.3	18.3	18.6	17.8
mGLR + SC-CI	40.2	40.4	40.0	31.6	32.5	31.4	24.0	24.5	23.3	18.9	19.2	18.4
$N(1.5, 0.2^2)$												
mSPRT	41.0	44.7	41.0	22.5	26.5	22.5	15.8	18.3	15.9	13.6	14.8	13.6
mGLR	24.9	25.4	25.0	20.0	21.1	19.9	15.9	17.1	15.9	13.6	14.7	13.6
mSPRT + SC-Standard	23.8	24.2	23.8	21.3	24.5	21.3	15.7	18.2	15.7	13.6	14.7	13.5
mSPRT + SC-MLE	13.0	12.9	13.1	12.8	12.7	12.8	12.4	12.5	12.4	12.0	12.2	12.0
mSPRT + SC-CI	23.1	20.6	23.1	21.0	20.5	21.0	15.7	17.6	15.7	13.6	14.6	13.5
mGLR + SC-Standard	20.9	21.3	20.9	18.9	19.8	18.9	15.7	17.0	15.8	13.6	14.6	13.5
mGLR + SC-MLE	12.9	12.8	13.0	12.7	12.7	12.7	12.4	12.5	12.4	12.0	12.2	12.0
mGLR + SC-CI	20.8	20.5	20.8	18.8	19.6	18.8	15.7	16.9	15.7	13.6	14.6	13.5

Table A-9. Percentage of Correct Classification for Cutoffs = (-1, 0, 1) and Broad Item Bank with True Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	87.9	87.1	87.8	88.1	87.2	87.7	86.8	86.8	86.8	85.4	85.8	85.7
mGLR	87.8	87.0	87.8	87.7	86.9	87.4	86.7	86.7	86.8	85.4	85.4	85.5
mSPRT + SC-Standard	87.8	87.0	87.8	88.1	87.1	87.7	86.8	86.8	86.8	85.4	85.8	85.7
mSPRT + SC-MLE	86.4	83.6	86.3	86.5	83.8	86.1	85.7	83.8	85.6	84.5	83.3	84.7
mSPRT + SC-CI	88.0	86.8	87.9	88.1	86.9	87.7	86.8	86.7	86.8	85.4	85.8	85.7
mGLR + SC-Standard	87.7	86.9	87.8	87.7	86.8	87.4	86.7	86.7	86.8	85.4	85.4	85.5
mGLR + SC-MLE	86.4	83.6	86.3	86.3	83.7	85.9	85.6	83.9	85.6	84.5	83.0	84.5
mGLR + SC-CI	87.9	86.8	87.9	87.7	86.8	87.4	86.7	86.7	86.8	85.4	85.4	85.5
$N(-0.5, 0.2^2)$												
mSPRT	93.4	93.1	92.9	93.4	93.1	92.9	93.2	92.5	92.2	92.1	90.2	91.1
mGLR	93.4	92.9	92.9	93.4	92.9	92.8	92.9	92.3	92.0	92.0	90.1	91.1
mSPRT + SC-Standard	93.4	93.0	92.8	93.4	93.0	92.8	93.2	92.5	92.2	92.1	90.2	91.1
mSPRT + SC-MLE	90.6	88.8	90.2	90.6	88.9	90.2	90.7	89.1	90.0	90.3	88.3	89.5
mSPRT + SC-CI	93.4	92.7	92.8	93.4	92.7	92.8	93.2	92.5	92.2	92.1	90.2	91.1
mGLR + SC-Standard	93.4	92.8	92.8	93.4	92.8	92.8	92.9	92.3	92.0	92.0	90.1	91.1
mGLR + SC-MLE	90.6	88.8	90.2	90.6	89.0	90.2	90.5	88.9	89.9	90.2	88.3	89.5
mGLR + SC-CI	93.4	92.7	92.8	93.4	92.8	92.8	92.9	92.3	92.0	92.0	90.1	91.1
$N(1.5, 0.2^2)$												
mSPRT	97.8	97.6	97.7	97.9	97.6	97.8	97.2	97.4	97.3	96.5	96.9	96.6
mGLR	97.9	97.6	97.8	97.8	97.7	97.7	97.3	97.5	97.4	96.5	96.9	96.6
mSPRT + SC-Standard	97.8	97.7	97.7	97.9	97.7	97.8	97.2	97.4	97.3	96.5	96.9	96.6
mSPRT + SC-MLE	97.9	97.7	97.9	97.9	97.7	97.9	97.3	97.5	97.4	96.6	96.8	96.8
mSPRT + SC-CI	97.9	97.7	97.8	97.9	97.7	97.8	97.2	97.4	97.3	96.5	96.9	96.6
mGLR + SC-Standard	97.9	97.7	97.8	97.8	97.7	97.7	97.3	97.5	97.4	96.5	96.9	96.6
mGLR + SC-MLE	97.9	97.7	97.9	97.8	97.7	97.8	97.4	97.5	97.5	96.6	96.8	96.8
mGLR + SC-CI	97.9	97.7	97.8	97.8	97.7	97.7	97.3	97.5	97.4	96.5	96.9	96.6

Table A-10. Average Test Length for Cutoffs = (-1, 0, 1) and Peaked Item Bank with True Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	45.2	45.0	44.8	29.7	30.4	29.1	21.1	22.1	20.4	16.2	17.1	16.0
mGLR	34.6	35.0	34.3	27.2	28.2	26.8	20.7	21.8	20.2	16.1	17.1	15.9
mSPRT + SC-Standard	41.3	41.0	40.8	29.3	29.9	28.8	21.0	22.0	20.3	16.2	17.1	16.0
mSPRT + SC-MLE	36.2	35.6	35.6	25.2	25.4	24.5	18.9	19.6	18.2	15.3	16.1	15.1
mSPRT + SC-CI	41.3	40.2	40.8	29.3	29.6	28.7	21.0	22.0	20.3	16.2	17.1	16.0
mGLR + SC-Standard	33.4	33.8	33.1	27.0	27.8	26.5	20.7	21.8	20.1	16.1	17.1	15.9
mGLR + SC-MLE	29.3	29.4	28.9	23.6	24.1	23.0	18.5	19.4	18.0	15.2	16.1	15.1
mGLR + SC-CI	33.5	33.8	33.2	26.9	27.7	26.5	20.6	21.8	20.1	16.1	17.1	15.9
$N(-0.5, 0.2^2)$												
mSPRT	46.6	45.7	45.2	28.0	28.4	27.2	19.5	20.5	18.8	15.1	16.2	15.0
mGLR	33.9	33.6	33.5	25.8	26.4	25.0	19.1	20.2	18.8	15.1	16.1	15.0
mSPRT + SC-Standard	46.4	45.4	45.0	27.9	28.4	27.2	19.5	20.5	18.8	15.1	16.2	15.0
mSPRT + SC-MLE	44.7	43.5	43.2	26.6	26.7	25.7	18.5	19.4	17.8	14.7	15.7	14.6
mSPRT + SC-CI	46.4	45.3	45.0	27.9	28.4	27.2	19.5	20.5	18.8	15.1	16.2	15.0
mGLR + SC-Standard	33.8	33.4	33.4	25.7	26.3	25.0	19.1	20.2	18.8	15.1	16.1	15.0
mGLR + SC-MLE	32.3	31.7	31.8	24.5	24.8	23.7	18.1	19.1	17.8	14.6	15.6	14.5
mGLR + SC-CI	33.8	33.4	33.4	25.7	26.3	25.0	19.1	20.2	18.8	15.1	16.1	15.0
$N(1.5, 0.2^2)$												
mSPRT	39.2	40.3	39.1	21.7	23.0	21.6	15.7	16.3	15.4	13.1	13.8	12.8
mGLR	24.2	24.3	24.1	19.1	19.4	18.9	15.5	16.0	15.3	13.1	13.8	12.9
mSPRT + SC-Standard	25.4	25.9	25.2	21.2	22.5	21.2	15.6	16.2	15.4	13.1	13.8	12.8
mSPRT + SC-MLE	13.2	13.2	13.1	12.9	13.0	12.8	12.5	12.5	12.3	12.1	12.2	11.9
mSPRT + SC-CI	24.1	21.6	24.0	21.2	21.1	21.1	15.6	16.2	15.4	13.1	13.8	12.8
mGLR + SC-Standard	21.4	21.7	21.3	18.7	19.0	18.5	15.5	15.9	15.2	13.1	13.8	12.9
mGLR + SC-MLE	13.1	13.1	13.0	12.9	13.0	12.7	12.5	12.5	12.2	12.1	12.2	11.8
mGLR + SC-CI	21.4	21.3	21.4	18.7	19.0	18.4	15.5	15.9	15.2	13.1	13.8	12.9

Table A-11. Percentage of Correct Classification for Cutoffs = (-1, 0, 1) and Peaked Item Bank with True Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	90.4	90.7	90.1	90.1	90.4	89.7	89.0	89.2	88.7	88.6	86.3	88.5
mGLR	90.4	90.6	90.0	89.9	90.4	89.7	88.9	89.1	88.5	88.6	86.4	88.4
mSPRT + SC-Standard	90.5	90.6	90.2	90.1	90.3	89.7	89.0	89.2	88.7	88.6	86.3	88.5
mSPRT + SC-MLE	88.7	88.2	88.4	88.4	87.9	88.0	87.5	86.8	87.4	86.8	85.0	86.9
mSPRT + SC-CI	90.5	90.6	90.2	90.2	90.3	89.8	89.0	89.2	88.7	88.6	86.3	88.5
mGLR + SC-Standard	90.5	90.5	90.1	89.9	90.4	89.7	88.9	89.1	88.5	88.6	86.4	88.4
mGLR + SC-MLE	88.7	88.1	88.3	88.2	87.9	88.0	87.4	86.8	87.2	86.8	85.0	86.8
mGLR + SC-CI	90.5	90.5	90.1	89.9	90.4	89.7	88.9	89.1	88.5	88.6	86.4	88.4
$N(-0.5, 0.2^2)$												
mSPRT	96.6	95.7	96.5	96.4	95.2	96.3	95.0	93.8	94.6	92.2	92.0	92.5
mGLR	96.5	95.4	96.3	96.0	94.9	95.9	94.9	93.8	94.7	92.0	92.0	92.5
mSPRT + SC-Standard	96.5	95.5	96.4	96.3	95.2	96.2	95.0	93.8	94.6	92.2	92.0	92.5
mSPRT + SC-MLE	93.2	92.2	93.0	93.0	91.9	92.8	92.0	91.1	91.5	90.4	90.1	90.6
mSPRT + SC-CI	96.5	95.5	96.4	96.3	95.2	96.2	95.0	93.8	94.6	92.2	92.0	92.5
mGLR + SC-Standard	96.4	95.3	96.2	96.0	94.9	95.9	94.9	93.8	94.7	92.0	92.0	92.5
mGLR + SC-MLE	93.2	92.1	92.9	92.8	91.7	92.6	91.9	91.2	91.6	90.2	90.1	90.6
mGLR + SC-CI	96.4	95.3	96.2	96.0	94.9	95.9	94.9	93.8	94.7	92.0	92.0	92.5
$N(1.5, 0.2^2)$												
mSPRT	98.0	97.8	98.1	98.0	97.8	98.2	97.6	96.9	97.3	96.5	96.3	96.3
mGLR	98.0	97.8	98.1	98.2	97.7	98.3	97.6	97.0	97.2	96.4	96.3	96.2
mSPRT + SC-Standard	98.0	97.8	98.2	98.0	97.8	98.2	97.6	96.9	97.3	96.5	96.3	96.3
mSPRT + SC-MLE	98.2	98.4	98.4	98.2	98.4	98.4	97.6	97.2	97.3	96.6	96.4	96.4
mSPRT + SC-CI	98.0	97.8	98.2	98.0	97.8	98.2	97.6	96.9	97.3	96.5	96.3	96.3
mGLR + SC-Standard	98.0	97.8	98.2	98.2	97.7	98.3	97.6	97.0	97.2	96.4	96.3	96.2
mGLR + SC-MLE	98.2	98.4	98.4	98.2	98.3	98.3	97.6	97.3	97.2	96.5	96.4	96.3
mGLR + SC-CI	98.0	97.8	98.2	98.2	97.7	98.3	97.6	97.0	97.2	96.4	96.3	96.2

Table A-12. Average Test Length for Cutoffs = (-1, 0, 1) and Broad Item Bank with Estimated Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	46.8	47.2	46.6	31.5	32.6	31.1	22.3	22.5	21.3	17.0	17.2	15.9
mGLR	36.4	36.9	36.4	28.9	29.3	28.7	21.9	22.2	21.2	16.5	17.1	16.0
mSPRT + SC-Standard	41.8	41.6	41.6	30.9	31.8	30.5	22.2	22.4	21.2	17.0	17.2	15.9
mSPRT + SC-MLE	36.6	36.6	36.5	26.5	27.0	26.1	19.7	20.0	18.8	16.1	16.4	15.1
mSPRT + SC-CI	42.1	41.1	41.9	30.9	31.3	30.5	22.1	22.4	21.2	17.0	17.2	15.9
mGLR + SC-Standard	34.7	35.0	34.7	28.3	28.7	28.2	21.8	22.1	21.1	16.5	17.0	15.9
mGLR + SC-MLE	30.3	30.7	30.3	24.5	25.0	24.4	19.4	19.9	18.7	15.6	16.2	15.1
mGLR + SC-CI	34.8	35.0	34.8	28.3	28.6	28.2	21.8	22.1	21.1	16.5	17.0	15.9
$N(-0.5, 0.2^2)$												
mSPRT	49.8	49.5	49.6	33.2	32.4	31.7	22.8	23.1	21.7	17.4	18.1	16.6
mGLR	38.1	38.0	38.1	29.5	29.2	29.2	22.1	22.7	21.2	16.6	17.8	16.5
mSPRT + SC-Standard	49.3	49.1	49.1	33.0	32.2	31.6	22.8	23.1	21.7	17.4	18.1	16.6
mSPRT + SC-MLE	47.6	47.1	47.6	31.5	30.5	30.2	21.9	22.1	20.8	17.0	17.7	16.2
mSPRT + SC-CI	49.4	49.0	49.2	33.0	32.2	31.6	22.8	23.1	21.7	17.4	18.1	16.6
mGLR + SC-Standard	37.8	37.7	37.7	29.4	29.1	29.0	22.0	22.7	21.2	16.5	17.8	16.5
mGLR + SC-MLE	36.2	35.9	36.3	28.0	27.7	27.7	21.1	21.7	20.4	16.2	17.4	16.1
mGLR + SC-CI	37.8	37.7	37.7	29.4	29.1	29.0	22.0	22.7	21.2	16.5	17.8	16.5
$N(1.5, 0.2^2)$												
mSPRT	39.8	43.6	39.8	21.4	23.9	21.3	15.2	16.5	15.0	12.9	13.5	12.7
mGLR	24.2	24.4	24.1	19.0	19.6	18.8	15.1	16.0	14.9	12.9	13.5	12.7
mSPRT + SC-Standard	23.0	23.3	23.0	20.7	22.9	20.6	15.1	16.4	14.9	12.9	13.5	12.7
mSPRT + SC-MLE	12.8	12.8	12.7	12.7	12.7	12.6	12.3	12.4	12.2	12.0	12.0	11.7
mSPRT + SC-CI	23.6	20.3	23.5	20.5	20.2	20.5	15.1	16.2	14.9	12.9	13.5	12.7
mGLR + SC-Standard	20.5	20.6	20.4	18.4	18.8	18.2	15.0	15.9	14.8	12.9	13.5	12.7
mGLR + SC-MLE	12.8	12.8	12.7	12.6	12.6	12.5	12.3	12.4	12.2	12.0	12.0	11.7
mGLR + SC-CI	20.6	20.2	20.6	18.3	18.7	18.1	14.9	15.9	14.8	12.9	13.5	12.7

Table A-13. Percentage of Correct Classification for Cutoffs = (-1, 0, 1) and Broad Item Bank with Estimated Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	87.6	86.1	87.6	87.5	85.9	87.4	86.4	85.0	86.4	84.1	83.0	84.1
mGLR	87.6	86.1	87.5	87.6	85.7	87.7	86.4	84.9	86.3	84.0	82.8	84.1
mSPRT + SC-Standard	87.5	85.9	87.5	87.5	85.9	87.4	86.4	85.0	86.4	84.2	83.0	84.2
mSPRT + SC-MLE	84.1	82.9	84.0	84.1	82.9	83.9	83.5	82.5	83.5	82.5	81.7	82.4
mSPRT + SC-CI	87.6	85.8	87.6	87.5	85.8	87.4	86.4	85.0	86.4	84.2	83.0	84.2
mGLR + SC-Standard	87.5	85.9	87.4	87.6	85.7	87.7	86.4	84.9	86.3	84.1	82.8	84.2
mGLR + SC-MLE	84.1	83.0	83.9	84.3	82.9	84.2	83.5	82.4	83.4	82.4	81.5	82.4
mGLR + SC-CI	87.6	85.9	87.5	87.6	85.7	87.7	86.4	84.9	86.3	84.1	82.8	84.2
$N(-0.5, 0.2^2)$												
mSPRT	94.7	94.3	94.5	94.7	94.3	94.5	93.3	92.9	93.3	91.4	91.3	91.6
mGLR	94.7	94.2	94.5	94.2	93.8	94.3	93.1	92.8	93.0	91.5	91.2	91.6
mSPRT + SC-Standard	94.6	94.2	94.5	94.7	94.2	94.5	93.3	92.9	93.3	91.4	91.3	91.6
mSPRT + SC-MLE	91.3	90.4	91.8	91.3	90.4	91.8	90.6	90.0	91.3	89.7	89.1	90.4
mSPRT + SC-CI	94.7	94.1	94.5	94.7	94.1	94.5	93.3	92.9	93.3	91.4	91.3	91.6
mGLR + SC-Standard	94.6	94.1	94.5	94.2	93.7	94.3	93.1	92.8	93.0	91.5	91.2	91.6
mGLR + SC-MLE	91.3	90.4	91.8	91.0	90.5	91.8	90.5	89.9	91.1	90.0	89.1	90.5
mGLR + SC-CI	94.7	94.1	94.5	94.2	93.7	94.3	93.1	92.8	93.0	91.5	91.2	91.6
$N(1.5, 0.2^2)$												
mSPRT	97.1	97.0	97.0	97.1	97.0	97.0	96.7	96.5	96.3	96.3	96.2	95.6
mGLR	97.1	97.1	97.0	97.0	96.8	96.8	96.8	96.4	96.4	96.3	96.2	95.6
mSPRT + SC-Standard	97.2	97.0	97.2	97.2	97.1	97.2	96.8	96.5	96.5	96.3	96.2	95.7
mSPRT + SC-MLE	97.8	98.2	97.7	97.8	98.2	97.7	97.3	97.3	96.9	96.5	96.7	95.9
mSPRT + SC-CI	97.2	97.2	97.2	97.2	97.2	97.2	96.8	96.5	96.5	96.4	96.2	95.8
mGLR + SC-Standard	97.2	97.1	97.2	97.1	96.9	97.0	96.9	96.4	96.6	96.3	96.2	95.7
mGLR + SC-MLE	97.8	98.2	97.7	97.7	97.9	97.5	97.3	97.3	96.9	96.5	96.7	95.9
mGLR + SC-CI	97.2	97.2	97.2	97.1	96.9	97.0	96.9	96.4	96.6	96.4	96.2	95.8

Table A-14. Average Test Length for Cutoffs = (-1, 0, 1) and Peaked Item Bank with Estimated Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	44.3	44.3	43.6	28.5	28.9	27.5	19.8	20.5	19.7	15.7	16.2	15.5
mGLR	33.7	34.0	33.2	26.2	26.6	26.0	19.5	20.3	19.5	15.6	16.0	15.5
mSPRT + SC-Standard	40.4	40.3	39.8	28.2	28.6	27.2	19.7	20.5	19.7	15.7	16.2	15.5
mSPRT + SC-MLE	35.7	35.0	35.0	24.7	24.3	23.8	18.2	18.6	18.1	15.2	15.5	14.9
mSPRT + SC-CI	40.5	39.4	39.8	28.2	28.3	27.2	19.7	20.4	19.7	15.7	16.2	15.5
mGLR + SC-Standard	32.6	32.6	32.1	25.9	26.3	25.7	19.5	20.2	19.5	15.6	16.0	15.5
mGLR + SC-MLE	28.8	28.3	28.3	22.9	23.1	22.7	18.0	18.5	18.0	15.1	15.3	14.9
mGLR + SC-CI	32.7	32.5	32.2	25.9	26.3	25.7	19.5	20.2	19.5	15.6	16.0	15.5
$N(-0.5, 0.2^2)$												
mSPRT	46.9	45.1	45.0	28.0	27.5	26.6	19.5	19.5	18.4	15.6	16.0	15.1
mGLR	33.4	33.2	32.9	25.2	25.4	24.5	19.2	19.3	18.3	15.4	15.9	15.1
mSPRT + SC-Standard	46.6	44.9	44.8	28.0	27.4	26.4	19.5	19.4	18.4	15.6	16.0	15.1
mSPRT + SC-MLE	45.0	42.4	43.5	26.6	25.4	25.4	18.6	18.3	17.8	15.1	15.4	14.8
mSPRT + SC-CI	46.7	44.8	44.8	28.0	27.4	26.4	19.5	19.4	18.4	15.6	16.0	15.1
mGLR + SC-Standard	33.2	33.0	32.7	25.1	25.3	24.4	19.1	19.3	18.3	15.4	15.9	15.1
mGLR + SC-MLE	31.7	31.0	31.5	23.8	23.6	23.4	18.2	18.2	17.7	14.9	15.2	14.8
mGLR + SC-CI	33.2	33.0	32.7	25.1	25.3	24.4	19.1	19.3	18.3	15.4	15.9	15.1
$N(1.5, 0.2^2)$												
mSPRT	35.7	36.5	35.7	18.3	19.9	18.4	13.8	14.7	14.0	12.0	12.7	12.1
mGLR	21.9	22.0	22.1	17.0	17.8	17.1	13.8	14.4	14.0	12.0	12.7	12.1
mSPRT + SC-Standard	23.7	24.3	23.8	18.0	19.6	18.0	13.8	14.6	14.0	12.0	12.7	12.0
mSPRT + SC-MLE	12.7	12.5	12.7	12.5	12.2	12.4	12.0	11.9	12.0	11.6	11.6	11.6
mSPRT + SC-CI	22.5	20.1	22.6	18.0	18.9	18.0	13.8	14.6	14.0	12.0	12.7	12.0
mGLR + SC-Standard	19.6	19.9	19.7	16.7	17.5	16.8	13.8	14.4	14.0	12.0	12.7	12.0
mGLR + SC-MLE	12.6	12.3	12.6	12.4	12.2	12.4	12.0	11.9	12.0	11.6	11.6	11.6
mGLR + SC-CI	19.4	19.6	19.6	16.6	17.4	16.7	13.8	14.4	14.0	12.0	12.7	12.0

Table A-15. Percentage of Correct Classification for Cutoffs = (-1, 0, 1) and Peaked Item Bank with Estimated Parameters

	$\delta = 0.1$			$\delta = 0.2$			$\delta = 0.3$			$\delta = 0.4$		
	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W	FI-C	FI-E	FI-W
$N(0, 1^2)$												
mSPRT	88.7	88.6	88.8	88.8	87.9	88.5	87.2	86.1	87.9	86.7	85.3	86.7
mGLR	88.7	88.1	88.7	88.5	87.3	88.2	87.2	85.8	87.8	86.6	85.1	86.7
mSPRT + SC-Standard	88.7	88.4	88.8	88.8	87.9	88.5	87.2	86.1	87.9	86.7	85.3	86.7
mSPRT + SC-MLE	87.4	86.3	87.5	87.6	85.8	87.2	86.0	84.8	86.7	85.9	84.3	86.0
mSPRT + SC-CI	88.6	88.2	88.7	88.8	87.6	88.5	87.2	86.0	87.9	86.7	85.3	86.7
mGLR + SC-Standard	88.7	87.9	88.7	88.5	87.3	88.2	87.2	85.8	87.8	86.6	85.1	86.7
mGLR + SC-MLE	87.4	85.9	87.4	87.4	85.8	86.9	86.0	84.6	86.5	85.9	84.1	86.0
mGLR + SC-CI	88.6	87.8	88.6	88.5	87.3	88.2	87.2	85.8	87.8	86.6	85.1	86.7
$N(-0.5, 0.2^2)$												
mSPRT	96.3	96.1	96.3	96.2	96.0	96.2	95.2	94.8	94.9	93.7	92.5	94.5
mGLR	96.3	96.0	96.1	96.1	95.9	96.2	95.3	94.6	94.9	93.8	92.7	94.5
mSPRT + SC-Standard	96.1	96.1	96.2	96.0	96.0	96.1	95.1	94.8	94.9	93.7	92.5	94.5
mSPRT + SC-MLE	93.1	91.0	93.8	93.1	91.0	93.8	92.7	90.1	93.0	92.1	89.3	93.1
mSPRT + SC-CI	96.2	96.1	96.2	96.0	96.0	96.1	95.1	94.8	94.9	93.7	92.5	94.5
mGLR + SC-Standard	96.1	96.0	96.0	95.9	95.9	96.1	95.2	94.6	94.9	93.8	92.7	94.5
mGLR + SC-MLE	93.2	91.0	93.7	93.0	91.0	93.8	92.8	90.1	93.0	92.2	89.5	93.1
mGLR + SC-CI	96.2	96.0	96.0	95.9	95.9	96.1	95.2	94.6	94.9	93.8	92.7	94.5
$N(1.5, 0.2^2)$												
mSPRT	97.9	98.0	97.7	97.8	97.9	97.5	97.5	97.3	97.1	96.9	96.9	96.4
mGLR	97.8	98.0	97.5	97.6	97.9	97.3	97.5	97.3	97.0	96.8	96.9	96.4
mSPRT + SC-Standard	97.9	98.0	97.8	97.8	97.9	97.5	97.5	97.3	97.1	96.9	96.9	96.4
mSPRT + SC-MLE	98.1	98.4	98.2	98.1	98.4	98.1	97.5	97.8	97.3	96.9	97.3	96.4
mSPRT + SC-CI	97.9	98.0	97.7	97.8	97.9	97.5	97.5	97.3	97.1	96.9	96.9	96.4
mGLR + SC-Standard	97.8	98.0	97.6	97.6	97.9	97.3	97.5	97.3	97.0	96.8	96.9	96.4
mGLR + SC-MLE	98.1	98.4	98.1	97.9	98.4	97.9	97.5	97.8	97.2	96.8	97.3	96.4
mGLR + SC-CI	97.8	98.0	97.5	97.6	97.9	97.3	97.5	97.3	97.0	96.8	96.9	96.4