Multidimensional Mastery Testing with CAT

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Steven Warren Nydick

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Niels Waller

Thesis Advisor

December, 2013

Acknowledgements

This dissertation reflects more than a simple study expanded over more than 200 pages. This paper represents who I am but also how I have grown over the course of my studies, as a student, a scholar, a friend, and a collaborator. Many people have earned my gratitude for their contribution to my time in graduate school. The following is only an incomplete list of those I have met along the way.

To my advisor, Niels Waller, who pushed me to be a better writer and scholar, helped me when needed, and allowed me to explore my own interests.

To my colleague and committee member, Chun Wang, who had faith in my research and writing abilities and helped me accomplish more professionally in the little time that I have known her than I had in the rest of my graduate studies.

To my committee members, David Weiss and Sandy Weisberg, who (as advisors of mine at one time in graduate school) have supported my studies and provided encouraging words and more than enough time and flexibility over the years.

To my undergraduate mentor, Benjamin Lovett, who had faith in my abilities and helped me discover my own path in graduate school. More than anyone else, he improved my writing ability through detailed and thoughtful criticism.

To the Psychology Department staff, especially Lynn Burchett and Judy Peterson, who assisted with all of the necessary administrative tasks, pushed me to hand in all

of my paperwork, and helped me navigate the complexities of two Masters degrees and one PhD.

To my network of supportive friends, Caren Arbeit, Mayumi Baker, Ben Babcock, Peter Marks, Stefani Quam, Hannah Riederer, and Shelby Vanderberg, who provided many ears (and drinks) over the years to help me express many of the frustrations of graduate school and escape from those frustrations.

To my colleagues on the sixth floor of Elliot Hall, Ben Babcock, Leah Feuerstahler, Rick Geyer, Chris Hulme-Lowe, Jieun Lee, Chaitali Phadke, and Dong Seo, who helped me think through problems and discussed the many intricacies of psychometrics that no one else would discuss.

To my colleagues at the American Registry of Radiologic Technologists and my co-interns at ACT, who provided me with projects that inspired fruitful ideas (including those that led to my dissertation) and challenged me to apply my abstract knowledge to practical problems.

To Jeff Jones, who always listened to my incredibly long rants, usually followed by beer, a bucket of hot wings, and a British comedy, and helped edit (and shorten) everything that I wrote in graduate school.

To my father, sister, and the rest of my family, who have always supported and encouraged me through the endless years of school and probably assumed that I would stay in graduate school at least three lifetimes into the future.

To my wonderful girlfriend, Lian Hortensius, who, according to friends, has made me smile much more than I did in the past.

To my mom. I love you and miss you.

Abstract

Computerized mastery testing (CMT) is a subset of computerized adaptive testing (CAT) with the intent of assigning examinees to one of two, mutually exclusive, categories. Most mastery testing algorithms have been designed to classify examinees on either side of a cut-point in one dimension, but many psychological attributes are inherently multidimensional. Little psychometric work has generalized these unidimensional algorithms to multidimensional traits. When classifying examinees in multidimensional space, practitioners must choose a cut-point function that separates a mastery region from a non-mastery region. The possible cut-point functions include one in which a linear combination of ability across dimensions must exceed a threshold and one in which each ability must exceed a threshold irrespective of any other ability. Moreover, two major components of every classification test are choosing successive questions and determining when a classification decision should be made. One frequently used stopping rule in unidimensional mastery testing is the Sequential Probability Ratio Test (SPRT), in which a classification is made either when the log-likelihood test statistic is sufficiently large or when the maximum number of items has been reached. Due to inefficiencies in the SPRT, alternative algorithms have been proposed, such as the Generalized Likelihood Ratio (GLR), and the SPRT with Stochastic Curtailment (SCSPRT). The current study explores properties of unidimensional classification testing algorithms, generalizes unidimensional methods to multidimensional mastery tests, and then tests many of the multidimensional procedures. Most of the multidimensional algorithms yield relatively efficient and accurate multidimensional classifications. However, some multidimensional classification problems, such as classifying examinees with respect to a linear classification bound function, are more robust to poor choices in the item bank or adaptive

testing algorithms. Based on results from the main study in this thesis, a follow-up study is proposed to better combine sequential classification methods with those based on directly quantifying incorrect classifications. I conclude by discussing consequences of the results for practitioners in realistic mastery testing situations.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Forty-five states have already adopted the Common Core State Standards (CCSS; 2010) to implement the dictates demanded by No Child Left Behind (NCLB; 2008). As explained on their website, the standards promise to adequately: (1) prepare students for college and work; (2) train students to compete in the global marketplace; and (3) determine student proficiency based on evidence of success (CCSS; 2010). Two state-led groups have been awarded federal funds to design assessments measuring objectives outlined in the CCSS. Due to the quantity of examinees, the high cost of exam development and implementation, and the consequence of mismeasurement, these assessments should quickly and accurately measure student readiness and achievement. Addressing the concerns of test developers, adaptive testing procedures base item selection and test length on the needs of an assessment and the responses of examinees to already administered items. Due to purported accuracy and efficiency, one of the state-led groups, the Smarter Balanced Assessment Consortium (SBAC; 2013), will soon adopt computerized adaptive tests (CAT; e.g., Wainer, 2000; Weiss, 1982) in high stakes exams (e.g., Way, Twing, Camara, Sweeney, Lazar, & Mazzeo, 2010).

Many tests, such as those constructed by the SBAC, seek to track individual changes

in ability. Another broad set of tests classifies examinees into categories based on the estimated location of ability relative to pre-specified cut-points. The most basic of the latter task-type determines classification by comparing ability to the minimal ability required for demonstrating competence in a particular field (e.g., Kingsbury & Weiss, 1983; Welch & Frick, 1993). For example, medical professionals are expected to know and conform to the dictates of their discipline lest they provide inadequate care and endanger patients' lives. These threshold-ability tests are generally referred to as "mastery" or "certification" tests (Bejar, 1983). Computerized mastery testing (CMT) is a subset of CAT with the intent of assigning examinees to one of two, mutually exclusive categories: one representing mastery and the other indicating non-mastery. Unlike CATs designed for equiprecise measurement (e.g., Weiss, 1982), the procedures implemented in CMT aim only increase the accuracy of certification.

Psychometric models designed for classification are divided into two, general areas: latent class models and latent trait models. Latent class models, such as diagnostic classification models (DCM; Rupp & Templin, 2008), assume that reality consists of a constellation of discrete cognitive states. One determines classification in DCM by estimating the posterior probability of each examinee having the attributes required for mastery given responses to test items. Latent trait models, such as item response theory (IRT; e.g., Embretson & Reise, 2000), assume that each examinee can be represented as a point in $\mathbb{R}^K$, where K is the number of dimensions underlying a series of test items. One determines classification in IRT by comparing the location of each examinee's ability vector in multidimensional space to some boundary curve separating the categories.

Both IRT and DCM can be used as the psychometric model underlying adaptive tests. All adaptive tests must include algorithms to determine which questions should be administered to each examinee and when enough information has been collected to end each test. During equiprecise CAT, questions are generally selected to provide as

much information as possible at the current ability estimate, and tests are generally stopped once that ability estimate has sufficiently stabilized. Conversely, during mastery tests, questions are generally selected to provide information about whether the examinee is on either side of the cut-point, and tests are generally stopped once that mastery decision has stabilized. These procedures will often result in drastically different tests. For example, imagine Einstein taking an introductory Physics equiprecise CAT. Because scarcely any questions are very difficult, the test would require many questions to differentiate his ability from other Physics professors. However, if the test were designed as a mastery test, only a few questions would be needed before a clear designation of "mastery" could be made.

Most researchers designing CMT algorithms using IRT models have assumed that only one trait underlies responses to test items (although see Glas & Vos, 2010; Seitz & Frey, 2013; and Spray, Abdel-fatah, Huang & Lau, 1997). Item selection algorithms for unidimensional, IRT-based mastery tests include selecting items by maximizing: (1) Fisher information at the classification bound (e.g., Eggen, 1999; Lin & Spray, 2000; Spray & Reckase, 1994); (2) Kullback-Leibler divergence at the classification bound (e.g., Eggen, 1999); and (3) the weighted log-odds ratio at the classification bound (e.g., Lin, 2011; Lin & Spray, 2000). Stopping rules for unidimensional, IRT-based mastery tests fall into two general categories: (1) Bayesian decision rules (e.g., Lewis & Shehan, 1990; Rudner, 2009); and (2) sequential decision theory (e.g., Bartroff, Finkelman, & Lai, 2008; Eggen, 1999; Finkelman, 2010; Thompson, 2009). The Bayesian decision approach to mastery testing determines, after each step, the posterior expected loss given prior information, classification proportions, and a set of responses. A test is then terminated if the expected loss for making a specific classification/decision is sufficiently small. In contrast, sequential decision theory algorithms are generally based off of Wald's (1945; 1947) Sequential Probability Ratio Test (SPRT), which uses a

likelihood ratio test statistic to determine when enough independent and identically distributed (i.i.d.) data have been collected to choose between one of two simple hypotheses. For unidimensional IRT-based CMT algorithms, these simple hypotheses are usually chosen to be specific ability values slightly within each category (e.g., Reckase, 1983). Then, after administering an item to an examinee, the SPRT must decide whether that examinee should be classified in the lower category, the examinee should be classified in the upper category, or the examinee should be administered another item. Primary justification for using the SPRT in mastery tests is attributed to the Wald-Wolfowitz theorem: given two simple hypotheses, "of all tests with the same power the sequential probability ratio test requires on the average fewest observations" (Wald & Wolfowitz, 1948). Because conditions underlying SPRT optimality do not generally apply to CMTs (see Nydick, 2012, for criticisms of the SPRT as applied to CMTs), researchers have proposed extensions of/alternatives to the simple SPRT, including the Generalized Likelihood Ratio Test (GLR; Bartroff, Finkelman, & Lai, 2008; Thompson, 2009), the SPRT with Stochastic Curtailment (SCSPRT; Finkelman, 2008a), and the SPRT with Predictive Power (Finkelman, 2010). The latter two methods, based off of probabilistic stopping rules taken from the clinical trials literature (Lan, Simon, & Halperin, 1982), determine whether to terminate an exam based, in part, on information from the remaining items in the bank.

Certification tests often aim to assess a composite of dimensions, but those tests generally provide a total score assumed to capture relevant information about that composite. For example, radiographers must know physics (how to use equipment), medicine (how to find the the appropriate anatomical area in a picture), patient care (how to make patients feel comfortable), etc. All of these dimensions are probably correlated to some degree, yet they are distinct enough to result in separate types of questions. Unfortunately, the criterion used to make a decision about the examinee's certification

is generally based on the total test score (Interpreting, 2003, p. 23). To classify an examinee along two dimensions using unidimensional IRT, one must make two separate decisions or, worse, pretend that those dimensions are perfectly correlated (although see Seitz & Frey and Spray et al., 1997, for alternative conceptions of "unidimensional algorithms" as applied to a multidimensional classification task). Multidimensional item response theory models are more flexible in accounting for relationships between these underlying dimensions unless those dimensions are highly correlated (Ackerman, 1989).

The purpose of this thesis is to extend the theory of unidimensional IRT-based computerized mastery testing to multidimensional IRT models and then to compare the proposed algorithms in a large simulation study. The remainder of this thesis is organized as follows. In Chapter 2, I explain the unidimensional three-parameter IRT model and review the most commonly used item selection algorithms and stopping rules in computerized mastery testing. In Chapter 3, I show how the optimal method of selecting items for unidimensional CMT depends on the IRT model used and the location of true ability relative to the classification bound. In Chapter 4, I describe the most commonly used multidimensional IRT models and extend unidimensional conceptualizations of mastery (including the item selection algorithms and stopping rules) to multiple dimensions. In Chapter 5, I outline a simulation study designed to compare many of the proposed multidimensional CTM algorithms in realistic testing situations. In Chapter 6, I summarize results from the simulation, and in Chapter 7, I draw conclusions from the simulation and propose future directions.

Chapter 2

Unidimensional Algorithms

In this chapter, I briefly outline the unidimensional, binary IRT model and explain currently used item selection algorithms and stopping rules for unidimensional adaptive mastery testing.

2.1 Unidimensional IRT and Mastery Testing

Item response theory (IRT) is a mathematical model that describes the relationship between responses to test items and examinee ability. The most popular IRT model remains the unidimensional, binary, three-parameter logistic model (3PL; Birnbaum, 1968) or simplifications thereof. Specifically, let θ represent the continuous latent variable underlying examinee responses to test items, assume that responses are conditionally independent[1] given a fixed $\theta = \theta_i$ (where i indexes examinees), and allow responses to have two possible scores, 0 and 1. Then, according to the 3PL, the probability of examinee i correctly responding to item j (i.e., getting a score of 1 on the item) is defined by the following item response function (IRF):

[1]For adaptive tests, "conditionally independent" is a more appropriate assumption than "locally independent" due to item selection dependencies (e.g., Mislevy & Chang, 2000).

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

where b_j denotes the inflection point of the IRF, a_j is proportional to the slope of the IRF at its inflection point, c_j indicates the lower asymptote, and D is a scaling constant usually specified to be either 1 (for the logistic metric) or 1.702[2] (for the normal-ogive metric), although alternative numbers for D have been proposed[3]. Because D is a scaling constant that does not affect model fit, I will let $D = 1$ for clarity. Common restrictions of the 3PL for binary items include: (1) eliminating the lower-asymptote parameter across all items, which results in a two-parameter model (2PL) specified by the following IRF:

$$p_j(\theta_i) = \Pr(Y_{ij} = 1|\theta_i, a_j, b_j) = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]}; \tag{2.2}$$

and (2) restricting the slope parameters across items to be identical, which results in a one-parameter (1PL) model, and can be written with the following IRF:

$$p_j(\theta_i) = \Pr(Y_{ij} = 1|\theta_i, b_j) = \frac{1}{1 + \exp[-a(\theta_i - b_j)]}. \tag{2.3}$$

Unidimensional models assume that a single ability underlies responses to all items on a test or in an item bank. Therefore, unidimensional mastery can be defined as a range of values on this latent dimension separated by a cut-score, θ_0. For examinee i, the correct mastery decision depends on the location of θ_i relative to θ_0. If $\theta_i > \theta_0$, the examinee should be classified as a master and any other decision is a Type II error. Conversely, if $\theta_i < \theta_0$, then the examinee should be classified as a failure and any other

[2] $D = 1.702$ minimizes the maximum difference between the normal and logistic cumulative distribution functions (Camilli, 1994).

[3] $D = 1.749$ minimizes the KL-divergence between the normal and logistic densities assuming the normal distribution is true (Savelli, 2006).

decision is a Type I error (Finkelman, 2008a). Unidimensional item selection algorithms and stopping rules were derived to result in the shortest tests conditional on pre-specified Type I and Type II error rates. In the next two sections, I briefly outline each of the commonly used item stopping rules and item selection algorithms in unidimensional CMT. Because efficient CMT item selection algorithms depend on the stopping rule, I first describe commonly used stopping rules and then explain how those classification criteria inform item selection decisions.

2.2 Unidimensional Stopping Rules

Many of the stopping rules in computerized mastery testing are modifications of Wald's Sequential Probability Ratio Test (SPRT; Wald, 1947). Therefore, I briefly outline the SPRT as applied to CMT and then review CMT-based modifications of the SPRT designed to circumvent its shortcomings.

2.2.1 The Sequential Probability Ratio Test

The classic stopping rule in CMT, the SPRT algorithm (e.g., Eggen, 1999; Reckase, 1983; Spray & Reckase, 1996), simplifies the classification task. Assume that a test administrator must classify examinees into one of two categories separated by a cut-point. Let θ_0 denote this a priori selected ability value separating true failures from true masters. Then point hypotheses can be specified as

$$\mathrm{H}_0 : \theta_i = \theta_0 - \delta$$

$$\mathrm{H}_1 : \theta_i = \theta_0 + \delta$$

where δ is a small constant putting H_0 slightly inside of the failure region and H_1 slightly

inside of the mastery region.

The purpose of any stopping rule in CMT is to determine whether an examinee should be classified as a master, a non-master, or be administered another item. To make one of these three decisions, the SPRT compares the likelihood ratio test statistic to appropriate critical values. As an example of how the likelihood ratio statistic might be applied, let responses be conditionally independent and follow the unidimensional, binary, item response function defined in Equation (2.1). Then the log-likelihood for a single examinee given a particular response pattern, $\mathbf{y}_{i,J} = [y_{i1}, y_{i2}, \ldots, y_{iJ}]^T$, is

$$\log[L(\theta|\mathbf{y}_{i,J})] = \sum_{j=1}^{J} \Big[y_{ij} \log[p_j(\theta)] + (1 - y_{ij}) \log[1 - p_j(\theta)] \Big] \tag{2.4}$$

with $p_j(\theta)$ defined in Equation (2.1). If $\mathrm{H}_0 : \theta_l = \theta_0 - \delta$ and $\mathrm{H}_1 : \theta_u = \theta_0 + \delta$, then the log-likelihood ratio of examinee i manifesting θ_u relative to θ_l is

$$C_{i,j} = \log \Big[\mathrm{LR}(\theta_u, \theta_l|\mathbf{y}_{i,j}) \Big] = \log \left[\frac{L(\theta_u|\mathbf{y}_{i,j})}{L(\theta_l|\mathbf{y}_{i,j})} \right] = \log \Big[L(\theta_u|\mathbf{y}_{i,j}) \Big] - \log \Big[L(\theta_l|\mathbf{y}_{i,j}) \Big]. \tag{2.5}$$

When Equation (2.5) is a large, positive number, then there is sizable evidence that θ_u generated the particular response pattern, $\mathbf{y}_{i,j}$. Conversely, when Equation (2.5) is a large, negative number, there is considerable evidence supporting θ_l.

Justification for using a likelihood ratio test statistic when testing simple hypotheses is due to the Neyman-Pearson lemma (Casella & Berger, 2001, p. 366). According to the Neyman-Pearson lemma, for a fixed sample size, N, and conditional on a particular Type I error rate, α, the uniformly most powerful (UMP) test rejects H_0 only contingent on the size of the likelihood ratio test statistic. Likelihood ratio-based test statistics are also optimal in the case of optional stopping, as proved in the

Wald-Wolfowitz theorem (Wald & Wolfowitz, 1948). Specifically, let $Y_1, Y_2, \ldots$ be a (possibly infinite) independent and identically distributed (i.i.d.) sample from common density f with unknown parameter vector $\boldsymbol{\theta}$ ($\dim(\boldsymbol{\theta}) \geq 1$). Then assuming a pair of simple hypotheses, $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ versus $\mathrm{H}_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_2$, and pre-specified critical values, A and B, where $0 < A < B < \infty$, a rule that stops sampling when $N = \inf\left\{n \geq 1 : \prod_{i=1}^{n}\left[\frac{f(y_i|\boldsymbol{\theta}_1)}{f(y_i|\boldsymbol{\theta}_2)}\right] \leq A \quad \text{or} \quad \prod_{i=1}^{n}\left[\frac{f(y_i|\boldsymbol{\theta}_1)}{f(y_i|\boldsymbol{\theta}_2)}\right] \geq B\right\}$ is optimal (i.e., minimizes the expected sample size under both H_0 and H_1) in the set of all tests with the same Type I and Type II error rates (Lai, 1997). Using the log-likelihood test statistic (rather than the likelihood) and given specific α (Type I error rate) and β (Type II error rate) levels, Wald (1947) recommended choosing $C_l = \log[A] = \log\left[\frac{\beta}{1-\alpha}\right]$ as the critical value separating non-mastery from uncertainty and $C_u = \log[B] = \log\left[\frac{1-\beta}{\alpha}\right]$ as the critical value separating mastery from uncertainty.

Modeled on sequential decision theory, psychometricians have designed a simple template for ending unidimensional mastery tests. After each item between the minimum number of items, $j_{\min}$, and the maximum number of items, $j_{\max}$, calculate $C_{i,j} = \log\left[\mathrm{LR}(\theta_u, \theta_l|\mathbf{y}_{i,j})\right]$ as defined in Equation (2.5). If $C_{i,j} < C_l$, classify the examinee as a failure and terminate the test. If $C_{i,j} > C_u$, classify the examinee as a master and terminate the test. But if $C_l \leq C_{i,j} \leq C_u$, administer another item. Once $j = j_{\max}$, use a final critical value of $(C_l + C_u)/2$ (Finkelman, 2008a) to make a decision. Often, researchers set $\alpha = \beta$, so that $(C_l + C_u)/2 = 0$, but practitioners sometimes desire to avoid one type of error depending on the ultimate costs of misclassification.

Unfortunately, researchers have identified several limitations of the standard SPRT in adaptive mastery testing. First, although Wald and Wolfowitz (1948) proved optimality of the SPRT when testing simple hypotheses, the SPRT is inefficient relative to other procedures if $\theta_i \neq \theta_l$ and $\theta_i \neq \theta_u$ (Finkelman, 2008a). In light of this concern, the Generalized Likelihood Ratio (GLR; Bartroff, Finkelman, & Lai, 2008; Thompson,

2009, 2010) was proposed as a simple modification of the SPRT that tests composite hypotheses. Second, the SPRT controls the error rate for infinitely long experiments under certain conditions, but every CAT must be terminated after a maximum number of items. Finkelman (2003, 2008a) proposed several procedures that use the likelihood ratio test statistic to estimate the probability of examinee i switching categories by $j_{\max}$. In the next several sub-sections, I explore each of the common adjustments to the SPRT algorithm.

2.2.2 The Generalized Likelihood Ratio

The Generalized Likelihood Ratio (GLR) is a modification of the simple SPRT algorithm for testing composite hypotheses. To derive the GLR procedure, consider a general version of the simple hypotheses specified above,

$$\begin{aligned} \mathrm{H}_0 &: \theta_i \leq \theta_l = \theta_0 - \delta \\ \mathrm{H}_1 &: \theta_i \geq \theta_u = \theta_0 + \delta, \end{aligned}$$

where $\theta \in \mathbb{R}$ has an associated likelihood function $f(\mathbf{y}|\theta)$. Then using a generalized likelihood ratio test statistic with $L(\theta_1|\mathbf{y}) = \arg\max_{\theta>\theta_0}\{f(\mathbf{y}|\theta)\}$ in the numerator and $L(\theta_2|\mathbf{y}) = \arg\max_{\theta\leq\theta_0}\{f(\mathbf{y}|\theta)\}$ in the denominator often results in a uniformly most powerful (UMP) test (Casella & Berger, 1990, p. 368). Intuitively, the generalized likelihood approach compares $\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta\in\mathbb{R}} L(\theta|\mathbf{y})$ (where MLE stands for Maximum Likelihood Estimate) to the most likely value of the composite hypothesis to which $\hat{\theta}_{\mathrm{MLE}}$ does not belong.

Adopting generalized likelihood ratio statistics in sequential analyses, Lai (2001) wrote that "simulation studies and asymptotic analyses have shown that [the number

of items needed to make a decision using a GLR] is nearly optimal over a broad range of parameter values θ, performing almost as well as [a procedure] that assumes θ to be known" (p. 307). Due to its desirable characteristics, Bartroff, Finkelman, and Lai (2008) proposed adopting

$$G_{i,j} = \log\left[\text{GLR}(\theta_0|\mathbf{y}_{i,j})\right] = \log\left[L(\hat{\theta}|\mathbf{y}_{i,j})\right] - \log\left[L(\theta'|\mathbf{y}_{i,j})\right] \tag{2.6}$$

as an alternative to the simple likelihood ratio in CMT, where $\theta' = \theta_0 + \delta$ if $\hat{\theta} \leq \theta_0$ or $\theta' = \theta_-^{j_{\max}}$ if $\hat{\theta} > \theta_0$, and $\theta_-^{j_{\max}}$ is found via Monte Carlo simulation to yield appropriate α and β rates. The same procedure is used in GLR as in SPRT with slightly different hypotheses, test statistics, and critical values (which are also found via simulation). Contrary to Bartroff et al. (2008), who proposed complicated methods for determining the test statistic and critical values, Thompson (2009, 2010) suggested that the GLR be identical to "the fixed point SPRT, with the exception that θ_1 and θ_2 [in the generalized likelihood ratio test statistic] are allowed to vary" (Thompson, 2010, p. 5). Therefore, Thompson advised calculating

$$\log\left[\text{GLR}(\theta_u, \theta_l|\mathbf{y}_{i,j})\right] = \sup_{\theta_1 \geq \theta_u}\left(\log\left[L(\theta_1|\mathbf{y}_{i,j})\right]\right) - \sup_{\theta_2 \leq \theta_l}\left(\log\left[L(\theta_2|\mathbf{y}_{i,j})\right]\right), \tag{2.7}$$

and comparing the result to C_l and C_u as in the SPRT. Note that Equation (2.7) contrasts the MLE with the maximum of the likelihood in the hypothesis to which the MLE does not belong. Regardless of method, both GLR and SPRT compare some version of a likelihood ratio test statistic to critical values that are only based on the items already taken. Finkelman (2003, 2008a) proposed a supplementary stopping rule, based on the work of Lan, Simon, and Halpern (1982) from the clinical trials literature, that also considers the remaining set of items before making a decision. I next address

the various curtailed decision rules.

2.2.3 The SPRT with Stochastic Curtailment

A curtailed version of a sequential procedure (Eisenberg & Ghosh, 1980) makes decision $D_{i,\,j} = r$ for examinee i, with $j < j_{\max}$, if and only if for every $s \neq r$, decision $D_{i,\,j_{\max}} = s$ can not happen. In other words, the curtailment criterion results in a decision if and only if all other decisions are impossible by the maximum sample size. Because a curtailment criterion is usually difficult to obtain, researchers have modified curtailed stopping rules to make decision $D_{i,\,j} = r$ for examinee i, with $j < j_{\max}$, if and only if for every $s \neq r$, the probability of deciding $D_{i,\,j_{\max}} = s$ is below some probability threshold (Finkelman, 2008a, p. 453). As applied to mastery testing, the SPRT with Stochastic Curtailment (SCSPRT; Finkelman, 2003) classifies an examinee when the probability of the examinee being classified in the other category by $j_{\max}$ is small.

To derive the stochastically curtailed SPRT, let $D_{i,\,j_{\text{tmp}}}$ be the temporary decision after $j_{\text{tmp}} < j_{\max}$ items, and assume that an SPRT-based mastery decision has not been made by j_{tmp} items. Set $D_{j_{\text{tmp}}} = n$ (where n stands for "non-master") if $C_{i,\,j_{\text{tmp}}} < (C_l{+}C_u)/2$, and set $D_{i,\,j_{\text{tmp}}} = d_m$ (where m stands for "master") if $C_{i,\,j_{\text{tmp}}} > (C_l{+}C_u)/2$. Next, pick two error rates, $0 \leq \epsilon_1 \leq .5$ and $0 \leq \epsilon_2 \leq .5$. Finally, classify the examinee as a non-master if $\{C_{i,\,j_{\text{tmp}}} < C_l\}$ or $\{D_{j_{\text{tmp}}} = n$ and $\Pr(D_{i,\,j_{\max}} = n | C_{i,\,j_{\text{tmp}}}) \geq 1 - \epsilon_1\}$; alternatively, classify the examinee as a master if $\{C_{i,\,j_{\text{tmp}}} > C_u\}$ or $\{D_{j_{\text{tmp}}} = m$ and $\Pr(D_{i,\,j_{\max}} = m | C_{i,\,j_{\text{tmp}}}) \geq 1 - \epsilon_2\}$. Notice how Finkelman (2008a) defined four error rates. α and β are the specified Type I and Type II error rates for an examinee at a particular end of the indifference region given an infinitely long experiment. Conversely, ϵ_1 and ϵ_2 are the specified Type I and Type II error rates for an examinee classified in a particular category at the hypothetical end of the test. Unlike α and β, ϵ_1 and ϵ_2 refer to a decision made by the end of the test and not the true classification.

To determine the SCSPRT decision rule, one must derive the probability of switching categories by maximum test length. Finkelman (2008a) used a normal approximation to the log-likelihood function after $j_{\max}$ items conditional on $j_{\text{tmp}} < j_{\max}$ items already administered. Specifically, define $C_0 = (C_l + C_u)/2$. Then after $j_{\text{tmp}} < j_{\max}$ items,

$$\Pr_{\tilde{\theta}}(D_{i,\,j_{\max}} = n | C_{i,\,j_{\text{tmp}}}) = 1 - \Pr_{\tilde{\theta}}(D_{i,\,j_{\max}} = m | C_{i,\,j_{\text{tmp}}}) \approx \Phi\left(\frac{C_0 - \mathbb{E}_{\tilde{\theta}}(C_{i,\,j_{\max}} | C_{i,\,j_{\text{tmp}}})}{\sqrt{\text{Var}_{\tilde{\theta}}(C_{i,\,j_{\max}} | C_{i,\,j_{\text{tmp}}})}}\right) \tag{2.8}$$

where

$$\mathbb{E}_{\tilde{\theta}}(C_{i,\,j_{\max}} | C_{i,\,j_{\text{tmp}}}) = C_{i,\,j_{\text{tmp}}} + \sum_{j=j_{\text{tmp}}+1}^{j_{\max}} \mathbb{E}_{\tilde{\theta}}\left(\log\left[\frac{L(\theta_u | Y_{ij})}{L(\theta_l | Y_{ij})}\right]\right), \tag{2.9}$$

$$\text{Var}_{\tilde{\theta}}(C_{i,\,j_{\max}} | C_{i,\,j_{\text{tmp}}}) = \sum_{j=j_{\text{tmp}}+1}^{j_{\max}} \text{Var}_{\tilde{\theta}}\left(\log\left[\frac{L(\theta_u | Y_{ij})}{L(\theta_l | Y_{ij})}\right]\right), \tag{2.10}$$

$\tilde{\theta}$ is the assumed ability under which the expectation/variance are evaluated, and $\Phi(\cdot)$ is the CDF of a standard normal distribution. One can straightforwardly calculate these probabilities as long as the remaining $j_{\text{tmp}} + 1$ to $j_{\max}$ are known in advance (or can be guessed) and if $j_{\text{tmp}} + 1 << j_{\max}$ for the Central Limit Theorem to apply (e.g., Finkelman, 2008a, p. 450). Non-nested likelihood ratio test statistics generally use an asymptotic normal distribution rather than a χ^2 distribution (Vuong, 1989).

Finkelman (2008a) proved that under mild sequential conditions, the SCSPRT replacing a generalized version[4] of the SPRT is weakly admissible[5]. Intuitively, Finkelman

[4] A generalized version of the SPRT is identical to the fixed-point SPRT with (potentially) step-dependent critical values. See Eisenberg, Gosh, and Simons (1976; as cited in Finkelman, 2008).

[5] A decision method is weakly admissible if no alternative method exists with at most as large error rates (with one of those error rates strictly smaller) and an almost-surely smaller number of sequential steps. See Eisenberg and Simons (1978; as cited in Finkelman, 2008).

(2008a) explained that weak admissibility of the SCSPRT results from stochastic curtailment "[eliminating] the use of all 'wasted' items, that is, all items that cannot affect the classification decision" (p. 455). Unfortunately, the Expectation and Variance in the Equations (4.15)–(4.16) are taken with respect to a particular θ: the closest endpoint of the indifference region, the current ability estimate, or (as Finkelman, 2008a, 2010, recommended) the endpoint of a confidence interval closest to the classification bound. Finkelman (2010) proposed a less ad hoc approach by weighting the conditional probability estimate on the distribution of θ_i after j_{tmp} items.

2.2.4 The SPRT with Predictive Power

Finkelman (2010) proposed several modifications of the stocastically curtailed SPRT stopping rule for computerized adaptive mastery tests. One option (termed "MLE formation") takes the expectation and variance of the likelihood ratio test statistic with respect to $\tilde{\theta} = \hat{\theta}_{\text{MLE}}$ rather than $\tilde{\theta} = \theta_l$ or $\tilde{\theta} = \theta_u$ once estimates of $\hat{\theta}_{\text{MLE}}$ stabilize. A second option (termed "confidence interval formation") is identical to the "MLE formation" but uses a confidence interval endpoint in the likelihood ratio test statistic rather than the MLE itself. The final recommendation of Finkelman (2010) (termed "predictive power") weights the SCSPRT by the posterior distribution of θ_i after j_{tmp} items. Specifically, let $\pi(\theta)$ be the prior distribution of θ. Then the posterior distribution of θ_i after j_{tmp} items can be written

$$\pi(\theta|\mathbf{y}_{i,\,j_{\text{tmp}}}) = \frac{\pi(\theta)L(\theta|\mathbf{y}_{i,\,j_{\text{tmp}}})}{\int_{\Theta}\pi(\theta)L(\theta|\mathbf{y}_{i,\,j_{\text{tmp}}})d\theta}, \tag{2.11}$$

where Θ is the set of all θ, $\mathbf{y}_{i,\,j_{\text{tmp}}}$ is the response pattern of examinee i after j_{tmp} items, $L(\theta|\mathbf{y}_{i,\,j_{\text{tmp}}})$ is the likelihood function given response pattern $\mathbf{y}_{i,\,j_{\text{tmp}}}$, and, by definition, the integral of a function $f(\theta)$, $\int_{\Theta} f(\theta)d\theta$, is taken over all $\theta \in \Theta$. Then the PPSPRT

can be defined as

$$\Pr_{\Theta}(D_{i,\,j_{\max}} = n | C_{i,\,j_{\text{tmp}}}) = \int_{\Theta} \pi(\theta | \mathbf{y}_{i,\,j_{\text{tmp}}}) \Pr_{\theta}(D_{i,\,j_{\max}} = n | C_{i,\,j_{\text{tmp}}}) d\theta, \tag{2.12}$$

where $\Pr_{\Theta}(D_{i,\,j_{\max}} = n | C_{i,\,j_{\text{tmp}}})$ is the expected SCSPRT criterion over Θ. If defining the loss in making a classification decision as

$$\text{Loss} = 100 \times I_W + J, \tag{2.13}$$

where I_W is an indicator function for incorrect classification and J is the number of items given to an examinee, then the PPSPRT and $\tilde{\theta} = \hat{\theta}_{\text{MLE}}$ methods resulted in the lowest average loss across all conditions (Finkelman, 2010). Therefore, using a PPSPRT stopping rule for mastery tests appears to result in a reasonable tradeoff between average test length and classification accuracy. Although SPRT-based stopping rules are increasingly used in CMT research, an alternative branch of CMT stopping rules are based on Bayesian decision theory.

2.2.5 Bayesian Decision Rules

An alternative set of stopping rules in CMT is based on Bayesian decision theory (e.g., Lewis & Shehan, 1990; Vos, 1999, 2002). Rather than adopting the likelihood ratio test statistic, Bayesian decision rules combine information from the posterior distribution of θ_i with specified costs in making decisions and administering items. Specifically, let Θ_m represent the set of masters, such that $\Theta_n = \Theta_m^c$ symbolizes the set of non-masters. Assuming that θ is a continuous random variable, then

$$\pi(m|\mathbf{y}_{i,\,j_{\text{tmp}}}) = 1 - \pi(n|\mathbf{y}_{i,\,j_{\text{tmp}}}) = \int_{\Theta_m} \pi(\theta|\mathbf{y}_{i,\,j_{\text{tmp}}})d\theta \tag{2.14}$$

is the posterior probability of mastery given the first j_{tmp} items. Given the posterior probability of mastery, one can then make a decision after determining: (1) states of nature, (2) possible actions, (3) loss functions, and (4) decision rules/principles. As an example of Bayesian decision theory, Lewis and Sheehan (1990) proposed a simple procedure. Define a mastery test with boundary point, θ_0, and possible states of nature, $\Theta = \{\Theta_n, \Theta_m\}$. Then, after item j is administered to examinee i, one can either fail the examinee ($\hat{\theta}_i \in \Theta_n$, where $\hat{\theta}_i$ is the MLE of θ_i), pass the examinee ($\hat{\theta}_i \in \Theta_m$), or administer another item. Each decision incurs a pre-specified cost. Let η_1 be the cost of passing an examinee who should not pass the test, η_2 be the cost of failing an examinee who should pass the test, and κ be the cost of administering an additional item. Using Equation (2.14), the expected loss of passing the examinee after item j can be written as

$$\mathbb{E}_\theta[L(\theta, m)|\mathbf{y}_{i,\,j}] = j\kappa + \eta_1 \cdot \big(1 - \pi(m|\mathbf{y}_{i,\,j})\big), \tag{2.15}$$

and the expected loss of failing the examinee after item j can be written as

$$\mathbb{E}_\theta[L(\theta, n)|\mathbf{y}_{i,\,j}] = j\kappa + \eta_2 \cdot \pi(m|\mathbf{y}_{i,\,j}). \tag{2.16}$$

The expected loss of administering another item depends on its usefulness for ultimately making a decision, and as such, relies on the expected future loss of eventually failing or passing the examinee. Lewis and Sheehan (1990) defined the risk at stage j as the expected loss incurred by making the best decision at that stage. Assuming that the decision after $j_{\max}$ items must be a classification, then the risk at the final stage can be

written

$$R_{j_{\max}}(\theta|\mathbf{y}_{i,\,j_{\max}}) = \min\left[j_{\max}\kappa+\eta_1\cdot(1-\pi\big(m|\mathbf{y}_{i,\,j_{\max}}\big)),\ j_{\max}\kappa+\eta_2\cdot\pi(m|\mathbf{y}_{i,\,j_{\max}})\right]. \quad (2.17)$$

Equation (2.17) can then be iteratively used to find the expected loss for administering another item at any j_{tmp} before $j_{\max}$. Specifically, the expected loss for continuing to test at stage j_{tmp} can be expressed as a function of the risk at stage $j_{\text{tmp}+1}$,

$$\mathbb{E}_\theta[L(\theta,c)|\mathbf{y}_{i,\,j_{\text{tmp}}}] = p_{j_{\text{tmp}}+1}(\theta)\cdot R_{j_{\text{tmp}}+1}(\theta|[\mathbf{y}_{i,\,j_{\text{tmp}}},1]) + \big(1-p_{j_{\text{tmp}}+1}(\theta)\big)\cdot R_{j_{\text{tmp}}+1}(\theta|[\mathbf{y}_{i,\,j_{\text{tmp}}},0]), \quad (2.18)$$

where $p_{j_{\text{tmp}}+1}(\theta)$ is the probability of correctly responding to item $j_{\text{tmp}}+1$, as defined in Equation (2.1), so the risk at stage $j_{\text{tmp}} < j_{\max}$ becomes

$$R_{j_{\text{tmp}}}(\theta|\mathbf{y}_{i,\,j_{\text{tmp}}}) = \min\left[\mathbb{E}_\theta[L(\theta,m)|\mathbf{y}_{i,\,j_{\text{tmp}}}],\ \mathbb{E}_\theta[L(\theta,n)|\mathbf{y}_{i,\,j_{\text{tmp}}}],\ \mathbb{E}_\theta[L(\theta,c)|\mathbf{y}_{i,\,j_{\text{tmp}}}]\right]. \quad (2.19)$$

Therefore, the potential risk at stage j_{tmp} depends on the possible risk at stages $j_{\text{tmp}}+1,\ldots,j_{\max}$. Equation (2.19) terminates in Equation (2.17) because the final stage must result in a pass/fail decision. After each item, the algorithm would choose the path (pass, fail, continue) that minimizes the corresponding expected loss. Note that $p_{j_{\text{tmp}}+1}(\theta)$ and $\pi\big(m|\mathbf{y}_{i,\,j_{\text{tmp}}+1}\big)$ must also be iteratively defined based on the distribution of responses given the $j^{\text{th}}_{\text{tmp}}$ posterior distribution of θ. Lewis and Shehan (1990) found that using a Bayesian decision procedure in lieu of administering a fixed number of items reduced CMTs by approximately 50% with little loss in classification accuracy.

Another common loss function is to penalize the decision based on the true distance

away from the classification bound (e.g., Vos, 1999), and another common decision rule is to minimize the maximum (rather than expected) loss (e.g., Vos, 2002). Regardless of loss function or decision rule, many stopping rules depend on a hypothetical complete test. For unidimensional mastery testing, several possible algorithms are available to choose those future items.

2.3 Unidimensional Item Selection Algorithms

All adaptive testing algorithms require methods of selecting future items. The current section details common item selection algorithms in adaptive tests and then explains modifications of those algorithms for use in mastery testing.

2.3.1 Fisher Information Methods

Many item selection algorithms require determining the information gained by, and thus the benefit of, choosing one potential future item over another potential future item. Fisher information (FI; Lord, 1980) measures the curvature of the log-likelihood in a small area surrounding the maximum likelihood estimate and relates to the asymptotic variance of $\hat{\theta}$ given true θ (e.g., Frank, 2009). Fisher information for item j can be written as a function of true θ,

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

where $p_j(\theta)$ is defined in Equation (2.1) and

$$p'_j(\theta) = \frac{dp_j(\theta)}{d\theta} = \frac{(1-c_j)a_j \exp[a_j(\theta - b_j)]}{(1+\exp[a_j(\theta - b_j)])^2} \tag{2.21}$$

is the derivative of $p_j(\theta)$ with respect to θ. The most common Fisher information-based item selection algorithm administers items that maximize (2.20) at $\theta = \hat{\theta}_i$. Basic criticisms of the original method include: (1) the likelihood function occassionally does not have a finite maximum (Veerkamp & Berger, 1997), and (2) the MLE estimate is often highly variable toward the beginning of a CAT (Chang & Ying, 1996). Other criticisms pertain to classification testing. For instance, if any $c_j > 0$, then selecting items based on maximizing Fisher information at $\hat{\theta}_i$ is not optimal in determining whether $\theta_i \in \Theta_m$ (e.g., Chapter 3; Spray & Reckase, 1994; Wiberg, 2003).

A variant of the typical Fisher information-based item selection algorithm is to take a weighted average of the information function across Θ (e.g., Veerkamp & Berger, 1997),

$$\mathcal{I}_j(\theta|w_{ij}) = \int_\Theta w_{ij}\mathcal{I}_j(\theta)d\theta, \tag{2.22}$$

where w_{ij} is the weight function for examinee i used for item j. Some common weight functions include $w_{ij} = 1$ iff $\theta = \hat{\theta}_i$, $w_{ij} = L(\theta|\mathbf{y}_{i,\,j-1})$, or $w_{ij} = \pi(\theta|\mathbf{y}_{i,\,j-1})$ (as defined in Equation 2.11). Using $w_{ij} = 1$ iff $\theta = \hat{\theta}_i$ is equivalent to standard Fisher information, and the latter two weight functions account for uncertainty in the maximum likelihood estimate. Veerkamp and Berger (1997) found that selecting items by maximizing likelihood-weighted Fisher information results in slightly shorter average test lengths than selecting items by maximizing Fisher information at $\hat{\theta}_{\text{MLE}}$. Another common algorithm described by Veerkamp and Berger (1997) is to maximize the average Fisher information across an interval. Let $\hat{\theta}_i^L$ be the lower limit of the interval and $\hat{\theta}_i^R$ be the upper limit of the interval. Then $w_{ij} = 1$ if $\theta \in [\hat{\theta}_i^L, \hat{\theta}_i^R]$ would result in choosing

an item that maximizes the average information across that interval. Averaging information across an interval considers many potential ability estimates and, thus, results in a more robust algorithm (as shown on p. 213 of Veerkamp & Berger for extreme values of θ). One could also circumvent the limited knowledge of locally-defined $\hat{\theta}_{\text{MLE}}$ by deriving a more globally-defined objective function.

2.3.2 Kullback-Leibler Methods

Chang and Ying (1996) suggested basing item selection on Kullback-Leibler (KL) divergence rather than FI as the information metric. KL divergence (e.g., Kullback, 1959; Kullback & Leibler, 1951) relates to the expected loss when choosing an approximate model rather than the correct model. Let f be the true probability distribution of univariate random variable X, and let g be an alternative/approximate distribution of X. Then the KL divergence between f and g is defined to be

$$\text{KL}(f||g) = \mathbb{E}_f\left(\log\left[\frac{f(X)}{g(X)}\right]\right) = \int_{-\infty}^{\infty} f(x)\log\left[\frac{f(x)}{g(x)}\right]dx, \tag{2.23}$$

where $||$ stands for "distance" between distributions, $\text{KL}(f||g) \geq 0$, and the expectation is taken with respect to the true distribution, f. To derive a KL-based index for computerized adaptive tests, let θ_i be the true ability of examinee i. Then KL divergence for the j^{th} item can be defined

$$\text{KL}_j(\theta_i||\theta) = \mathbb{E}_{\theta_i}\log\left[\frac{L(\theta_i|Y_{ij})}{L(\theta|Y_{ij})}\right] = p_j(\theta_i)\log\left[\frac{p_j(\theta_i)}{p_j(\theta)}\right] + [1-p_j(\theta_i)]\log\left[\frac{1-p_j(\theta_i)}{1-p_j(\theta)}\right]. \tag{2.24}$$

As shown by Chang and Ying (1996), the curvature of the KL divergence function at a point is equal to Fisher information at that point. Therefore, KL divergence effectively

reduces to Fisher information if choosing θ to be close to θ_i. Chang and Ying (1996) recommended using KL divergence because "there is no requirement that θ_i be close to θ" (p. 218), unlike the more local character of Fisher information. Moreover, by using a likelihood ratio statistic, KL divergence is similar to the decision-making process of the SPRT. Several different KL divergence indices have been proposed for use in adaptive testing. Originally, the KL information index was defined as average KL divergence along a small interval,

$$\mathrm{KL}_j(\hat{\theta}_i) = \int_{\hat{\theta}_i - \delta_{ij}}^{\hat{\theta}_i + \delta_{ij}} \mathrm{KL}(\hat{theta_i}||\theta)d\theta, \tag{2.25}$$

where $\hat{\theta}_i$ is the MLE of θ_i before administering item j, and δ_{ij} is a function of the precision in the MLE. Chen, Ankenmann, and Chang (2000) also noted that, as in Fisher information, weight functions can be applied to KL divergence indices, resulting in

$$\mathrm{KL}_j(\hat{\theta}_i|w_{ij}) = \int_{\hat{\theta}_i - \delta_{ij}}^{\hat{\theta}_i + \delta_{ij}} w_{ij}\mathrm{KL}(\hat{\theta}_i||\theta)d\theta. \tag{2.26}$$

They further compared bias, MSE, and item overlap for various FI and KL criteria across CATs designed to estimate θ_i for each person. Only for extreme ability levels did KL information or weighted Fisher information improve over standard Fisher information in terms of bias, MSE, and item overlap early in a test. Moreover, as they wrote, "differences among all [item selection algorithms] with respect to BIAS, RMSE, SE, and item overlap were negligible for tests of more than 10 items" (p. 253, and see Cheng and Lio, 2000, for a partial replication of this study with nearly identical results).

All of the item selection algorithms heretofore discussed were derived to pinpoint an examinee's true ability. None of the algorithms as presented can be used to efficiently decide whether an examinee is in one of two broadly defined categories. In Chapter 3, I

show why each of the aforementioned algorithms results in inefficient CMTs. However, before explaining reasons for inefficiencies, I first describe alternatives to the typical item selection algorithms appropriate for unidimensional adaptive mastery tests.

2.3.3 Mastery Testing Methods

Many researchers have suggested modifications of the above algorithms for use in mastery testing. For instance, Eggen (1999) promoted selecting items by maximizing Fisher information at θ_0 rather than $\hat{\theta}_i$ or maximizing point-wise KL divergence (Equation 2.24) at $\mathrm{KL}_j(\theta_u||\theta_l)$. He found that maximizing Fisher information at the cut-point resulted in the shortest and most accurate tests, but selecting items to maximize Fisher information at $\hat{\theta}_i$ or KL divergence using $\mathrm{KL}_j(\theta_u||\theta_l)$ did not result in much performance decrement (although see Eggen, 2010 for a replication of Eggen, 1999 with slightly different results).

A common complaint in using point-wise KL divergence in mastery testing is the lack of symmetry between $\mathrm{KL}_j(\theta_u||\theta_l)$ and $\mathrm{KL}_j(\theta_l||\theta_u)$. Recall that KL divergence is defined as the expected log-likelihood ratio comparing the true model to an alternative model *with respect to the true model*. Therefore, when choosing items by maximizing $\mathrm{KL}_j(\theta_u||\theta_l)$, one implicitly assumes that every examinee is a master. Alternative mastery testing item selection algorithms have been developed that better consider the actual location of an examinee when selecting items. These alternative algorithms include the weighted log-odds ratio (LO; Lin & Spray, 2000) and mutual information (MI; Weissman, 2007). The weighted log-odds ratio selects items that maximize the expected log-odds at the ends of the indifference region,

$$\mathrm{LO}_j(\theta_u||\theta_l) = \sum_y \mathbb{E}\log\left(\left[\frac{p_j(\theta_u)}{p_j(\theta_l)}\right]^Y \div \left[\frac{1-p_j(\theta_u)}{1-p_j(\theta_l)}\right]^{1-Y}\right) \tag{2.27}$$

$$= \mathbb{E}(Y=1)\log\left[\frac{p_j(\theta_u)}{p_j(\theta_l)}\right] - [1-\mathbb{E}(Y=1)]\log\left[\frac{1-p_j(\theta_u)}{1-p_j(\theta_l)}\right], \tag{2.28}$$

where $\mathbb{E}(Y=1)$ is the classical difficulty of an item and can be calculated by integrating the probability of response for θ weighted on the density of θ across the examinee distribution[6].

Mutual information generalizes log-likelihood-based information criteria across multiple cut-points. Weissman (2007) proposed MI as a symmetric version of KL divergence. Let Θ_B be a discrete set describing the classification bound(s). In our case, $\Theta_B = \{\theta_l, \theta_u\}$. Then mutual information can be defined as

$$\begin{aligned}\mathrm{MI}_j(\Theta_B) &= \sum_y \sum_{\theta\in\Theta_B} f(y,\theta)\log\left[\frac{f(y,\theta)}{f(y)f(\theta)}\right] \\ &= \sum_y \sum_{\theta\in\Theta_B} \Pr_j(Y=y|\theta)\pi(\theta)\log\left[\frac{\Pr_j(Y=y|\theta)}{f(y)}\right],\end{aligned} \tag{2.29}$$

where $\Pr_j(Y=y|\theta)$ is the probability of $Y=y$ given a particular θ, $\pi(\theta)$ is the prior probability of θ, and $f(y)$ is the marginal probability of $Y=y$. Lin (2011) tested FI, KL, LO, and MI in several SPRT-based CMTs. He found that the weighted log-odds ratio resulted in the fewest number of items administered, and mutual information resulted in the most number of items administered. All of the algorithms had similar classification accuracies. In Chapter 4, I discuss generalizations of the item selection

[6]Lin and Spray, 2000 and Lin, 2011 take the expectation in Equation (2.27) with respect to the marginal distribution of θ to arrive at Equation (2.28). However, I found taking the expectation with respect to a single examinee's $\hat{\theta}_i$ to better reflect the associated SPRT stopping rule. The latter item selection rule will be expounded upon in later chapters.

and stopping rules to multidimensional adaptive tests. But first, I explain limitations of using certain item selection rules in adaptive mastery tests as a partial justification for deriving particular multidimensional mastery testing item selection algorithms.

Chapter 3

SPRT and Binary Response Models

A potential limitation of using the SPRT as a decision rule in unidimensional classification tests is due to the non-zero lower asymptote of the three-parameter logistic model. Spray and Reckase (1994) noticed that when using the 3PL, "selecting items to have maximum information at the examinee's true ability results in longer average test lengths" (p. 9) than selecting items to have maximum information at the cut-points, and "this result is quite dramatic for the lower [classification bound] and examinees above θ of .5" (p. 9). In other words, the SPRT is inefficient for high ability simulees when using the three-parameter logistic model and selecting items based on the maximum likelihood estimate. Spray and Reckase (1994) proposed a simple method of reducing the number of administered items in SPRT-based classification tests: select items to maximize information at the cut-point separating categories. However, the ideal item selection point depends on the true item and person parameters as well as the classification bound. Selecting items to maximize information at the classification bound is

only a coarse approximation of the most efficient item selection algorithm. By examining properties of IRT log-likelihood ratios, one can shed light on optimal methods of designing item banks, choosing item selection algorithms, and selecting classification criteria for adaptive tests. Because multidimensional IRT models are generalizations of the unidimensional functional form, many of these results should also apply to multidimensional adaptive tests. In the following sections, I present the effect of item and person parameters on the magnitude of the SPRT test statistic in two parts: first with mathematical evidence, and then, supporting mathematical conclusions with a small set of simulations.

3.1 Mathematical Considerations

In this section, I demonstrate how the SPRT-based test statistic changes as properties of the logistic model are altered. Thompson (2010), who used the 3PL in his simulations, wrote that "it is far easier to make a classification if the cut-score is in the extremes" and that "typically, only a few items might be needed to classify an examinee above a cut-score of -2.0 or below $+2.0$" (p. 9). Thompson's assertion is accurate when using the GLR as a stopping rule (which was the purpose of his paper) but not always when adopting the SPRT. Because his discussion includes research on both stopping rules, his statement of classification efficiency is not entirely correct. Only Spray and Reckase (1994) explicitly acknowledged that "the large difference in number of items for the high ability examinees [when selecting items at proximate estimates of θ rather than at the cut-point] is a result of the nonzero lower asymptote for the three parameter logistic model" (p. 7). I now briefly show why non-zero lower asymptotes affect the magnitude of a likelihood ratio test statistic in certain situations. The first part of this section focuses on properties of the classification bound. I then reverse the investigation

by examining optimal items given test attributes.

3.1.1 The SPRT Test Statistic and Classification Bounds

Consider a classification task involving one cut-point and a symmetric indifference region of size 2δ. Conceiving the likelihood ratio test statistic as a function of the classification bound, θ_0, Equation (2.5) for examinee i after a fixed set of J items can be written

$$\begin{aligned}\log\left[\mathrm{LR}(\theta_0+\delta,\theta_0-\delta|\mathbf{y}_i)\right] &= \log\left[\frac{L(\theta_0+\delta|\mathbf{y}_i)}{L(\theta_0-\delta|\mathbf{y}_i)}\right] \\ &= \sum_{j=1}^{J} y_{ij}\log\left[\frac{p_j(\theta_0+\delta)}{p_j(\theta_0-\delta)}\right] + \sum_{j=1}^{J}(1-y_{ij})\log\left[\frac{1-p_j(\theta_0+\delta)}{1-p_j(\theta_0-\delta)}\right]. \qquad (3.1)\end{aligned}$$

Spray and Reckase (1994) noticed that when $c_j > 0$ then $\lim_{\theta_0\to-\infty}\frac{p_j(\theta_0+\delta)}{p_j(\theta_0-\delta)} = 1$ and $\lim_{\theta_0\to-\infty}\frac{1-p_j(\theta_0+\delta)}{1-p_j(\theta_0-\delta)} = 1$. Therefore, when all of the pseudo-guessing parameters are greater than 0 and the classification bound is extremely negative (e.g., 4 or more standard deviations below the average θ_i), Equation (3.1) will be close to 0 regardless of an examinee's true ability. Note that the above situation is not realistic in practice, as it assumes the cut-point approaches negative infinity with a fixed set of item difficulties. Most practicable tests use cut-points well within the range of the item parameters. But if items are not strategically selected on a CAT, the limiting problems of the log-likelihood ratio can be approximated even with a well-designed item bank.

One can better understand the behavior of the log-likelihood ratio by studying its change in slope. Taking the first derivative of Equation (3.1) with respect to θ_0 results in

$$\frac{d\log\left[\mathrm{LR}(\theta_0+\delta,\theta_0-\delta|\mathbf{y}_i)\right]}{d\theta_0} = \sum_{j=1}^{J} a_j y_{ij}[p_j^{c_j}(\theta_0+\delta) - p_j^{c_j}(\theta_0-\delta)] - \sum_{j=1}^{J} a_j[p_j^{1}(\theta_0+\delta) - p_j^{1}(\theta_0-\delta)], \quad (3.2)$$

where $p_j^{c_j}(\theta_0) = \frac{\exp[a_j(\theta_0 - b_j)]}{c_j + \exp[a_j(\theta_0 - b_j)]}$ and $p_j^1(\theta_0) = \frac{\exp[a_j(\theta_0 - b_j)]}{1 + \exp[a_j(\theta_0 - b_j)]}$. If $c_j = 0$ for all $j \in \{1, \ldots, J\}$, then $p_j^{c_j}(\theta_0 + \delta) - p_j^{c_j}(\theta_0 - \delta) = 1 - 1 = 0$ for all items, so that

$$\frac{d \log \left[\text{LR}(\theta_0 + \delta, \theta_0 - \delta | \mathbf{y}_i) \right]}{d\theta_0} = -\sum_{j=1}^{J} a_j [p_j^1(\theta_0 + \delta) - p_j^1(\theta_0 - \delta)], \tag{3.3}$$

which does not depend on item responses. Importantly, the sign of Equation (3.3) is always negative for the 1PL and 2PL (unless $\delta = 0$), so that the log-likelihood ratio test statistic is monotonically decreasing as the location of the classification bound increases. The consequence of a monotonic log-likelihood ratio statistic can be explained with a simple example. Assume that an examinee has true $\theta_i = 2.0$ and is taking a classification test with two classification bounds: $\theta_{0_a} = 0$ and $\theta_{0_b} = 1.0$. Note that θ_i is much further from the $\theta_{0_a} = 0$ cut-point than the $\theta_{0,b} = 1.0$ cut-point. If $c_j = 0$ for all items on the exam, then the log-likelihood ratio test statistic will be larger comparing $\theta_{0_a} + \delta$ to $\theta_{0_a} - \delta$ than comparing $\theta_{0_b} + \delta$ to $\theta_{0_b} - \delta$, providing more evidence that examinee i is above $\theta_{0_a} = 0$ than $\theta_{0_b} = 1.0$.

Unfortunately, if any $c_j > 0$, then the log-likelihood ratio is not necessarily monotonic. To see the consequences of non-monotonicity for classification evidence, take expectations of Equation (3.2) conditional on θ_i and compare the outcome to zero, which results in

$$\begin{aligned}
\sum_{j=1}^{J} a_j [p_j^1(\theta_0 + \delta) - p_j^1(\theta_0 - \delta)] &\gtreqless \sum_{j=1}^{J} a_j p_j(\theta_i) [p_j^{c_j}(\theta_0 + \delta) - p_j^{c_j}(\theta_0 - \delta)], \\
&\gtreqless \sum_{j=1}^{J} a_j \left[\left(\frac{p_j(\theta_i)}{p_j(\theta_0 + \delta)} \right) p_j^1(\theta_0 + \delta) - \left(\frac{p_j(\theta_i)}{p_j(\theta_0 - \delta)} \right) p_j^1(\theta_0 - \delta) \right], \\
\sum_{j=1}^{J} a_j p_j^1(\theta_0 + \delta) \left[1 - \frac{p_j(\theta_i)}{p_j(\theta_0 + \delta)} \right] &\gtreqless \sum_{j=1}^{J} a_j p_j^1(\theta_0 - \delta) \left[1 - \frac{p_j(\theta_i)}{p_j(\theta_0 - \delta)} \right],
\end{aligned} \tag{3.4}$$

where $\gtreqless$ indicates the left side of Equation (3.4) will be greater than, equal to, or less

than the right side. Assume that a test contains one item with $a = 1$, $b = 0$, $c = .2$, a single examinee takes the test with true $\theta_i = 2.0$, and the SPRT stopping rule is used for classification with $\delta = .1$. Then each half of Equation (3.4) is displayed on the left side of Figure 3.1, and the full expected derivative is displayed on the right side of Figure 3.1 for various values of θ_0. When $\theta_0 < -0.94$, then $p^1(\theta_0 + \delta)\left[1 - \frac{p(\theta_i)}{p(\theta_0+\delta)}\right] < p^1(\theta_0 - \delta)\left[1 - \frac{p(\theta_i)}{p(\theta_0-\delta)}\right]$, but at approximately $\theta_0 = -0.94$, the curves cross, and then $p^1(\theta_0 + \delta)\left[1 - \frac{p(\theta_i)}{p(\theta_0+\delta)}\right] > p^1(\theta_0 - \delta)\left[1 - \frac{p(\theta_i)}{p(\theta_0-\delta)}\right]$. For these set of parameters, the strongest evidence for $\theta_i > \theta_0$ is when $\theta_0 \approx -0.94$ and not when $\theta_0 < -2.0$. In fact, when $\theta_0 = -4.0$, the expected log-likelihood ratio is approximately .013, whereas when $\theta_0 = -0.94$, the expected log-likelihood ratio is approximately .076, providing additional evidence that θ_i is in the upper category.

To better understand the consequences of item parameter values on classification evidence, one can analytically solve for the maximum of the log-likelihood function. Pretend that an examinee has correctly responded to a single item test from an item bank calibrated under the 3PL. To find the corresponding classification bound resulting in the strongest evidence for classification, construct the log-likelihood ratio assuming a correct response and given a specified indifference region,

$$f_1(\theta_0) = \log\left[\frac{p(\theta_0 + \delta)}{p(\theta_0 - \delta)}\right] = \log[p(\theta_0 + \delta)] - \log[p(\theta_0 - \delta)], \tag{3.5}$$

with $p(\theta)$ defined in Equation (2.1) and a, b, and c dependent on the particular item chosen, set the derivative of Equation (3.5) equal to 0, and solve for θ_0. As shown in Appendix A.1, one finds that Equation (3.5) is maximized when

$$\hat{\theta}_0 = \frac{\log(c)}{2a} + b. \tag{3.6}$$

Because $\log(c) < 0$ for $c \in (0, 1)$, the optimal classification bound for a correct item is

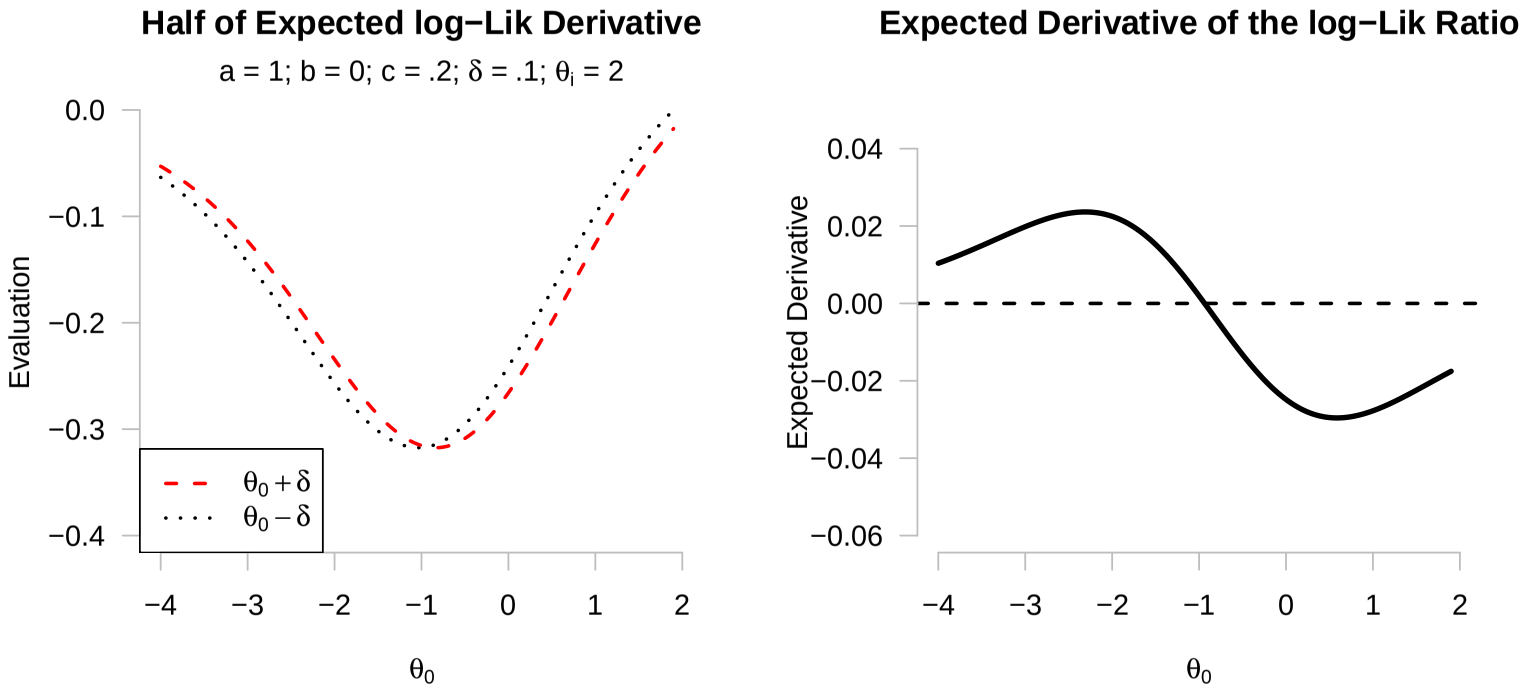

Figure 3.1: The expected derivative of the SPRT log-likelihood ratio for various values of θ_0. The left plot shows each half of the expected derivative of the log-likelihood ratio test statistic when $a = 1$, $b = 0$, $c = .2$, $\theta_i = 2.0$, $\delta = .1$, and θ_0 is varied from -4.0 to 2.0 as displayed in Equation (3.4). The right plot shows the full expected derivative as presented in Equation (3.2).

somewhat below the item difficulty and does not depend on the size of the indifference region. Necessarily, as $c \to 0$, then the right side of Equation (3.6) approaches $-\infty$, but as $c \to 1$ or $a \to \infty$, then the classification bound that maximizes the log-likelihood ratio approaches $\theta_0 = b$. Therefore, when $c > 0$, larger item discriminations do not mitigate the effect of a lower asymptote on the optimal classification bound.

Of course, one does not know a priori that an examinee will respond to an item in a particular way. One can instead find the classification bound that optimizes the expected log-likelihood ratio for a single item,

$$\begin{aligned} f_2(\theta_0) &= \mathbb{E}_{\theta_i}\left[\log\left[\mathrm{LR}(\theta_0+\delta,\theta_0-\delta|Y)\right]\right] \\ &= p(\theta_i)\log\left[\frac{p(\theta_0+\delta)}{p(\theta_0-\delta)}\right] + [1-p(\theta_i)]\log\left[\frac{1-p(\theta_0+\delta)}{1-p(\theta_0-\delta)}\right]. \end{aligned} \tag{3.7}$$

with all of the terms identical to those in Equation (3.5). As shown in Appendix A.2, the maximum of Equation (3.7) is found to be

$$\hat{\theta}_0 = \frac{\log(c)}{2a} + \theta_i - \frac{\log\left(\left[\exp[a(\theta_i-b)]\{c+1+\exp[a(\theta_i-b)]\} + \left(c^{1/2}\cosh[a\delta]\right)^2\right]^{1/2} + \left(c^{1/2}\cosh[a\delta]\right)\right)}{a}. \tag{3.8}$$

One immediately deduces several consequences of Equation (3.8). First, the classification bound that maximizes the expected log-likelihood ratio for a single item is monotonically increasing with respect to θ_i, a, b, c, and δ (assuming that a, c, $\delta > 0$). Second, holding a, b, c, and δ constant, the maximal classification bound for large θ_i,

$$\begin{aligned} \hat{\theta}_0 &\approx \frac{\log(c)}{2a} + \theta_i - \frac{\log\left(\left[\exp[2a(\theta_i-b)]\right]^{1/2}\right)}{a} \\ &= \frac{\log(c)}{2a} + \theta_i - \frac{a(\theta_i-b)}{a} = \frac{\log(c)}{2a} + b, \end{aligned} \tag{3.9}$$

is identical to the maximal classification bound assuming a correct response. Finally, as long as $c > 0$, Equation (3.7) has a finite maximum that is slightly below b.

3.1.2 The SPRT Test Statistic and Item Difficulties

Thus far, I have shown the effect of θ_0 on the expected SPRT test statistic given fixed person and item parameters. Practitioners generally fix θ_0 and determine the optimal item to administer based on statistical considerations. Instead of examining the expected log-likelihood ratio as a function of the classification bound, θ_0, one could instead let Equation (3.7) be a function of the item difficulty,

$$f_2(b) = p_j(\theta_i) \log \left[\frac{p_j(\theta_0 + \delta)}{p_j(\theta_0 - \delta)} \right] + [1 - p_j(\theta_i)] \log \left[\frac{1 - p_j(\theta_0 + \delta)}{1 - p_j(\theta_0 - \delta)} \right], \tag{3.10}$$

and then optimize Equation (3.10) with respect to b. If $c > 0$, then the maximum of Equation (3.10) is not analytically feasible. Setting $c = 0$, the optimal b-parameter is shown in Appendix A.3 to be

$$\hat{b} = \log \left[\frac{-\gamma + \sqrt{\gamma^2 - 4\omega\psi}}{2\omega} \right] / a, \tag{3.11}$$

where

$$-\gamma = 4(a\delta \cosh[a\delta] - \sinh[a\delta]) \exp[a(\theta_0 + \theta_i)], \tag{3.12}$$

$$2\omega = 4 \sinh[a\delta] \exp[a\theta_0] - 4a\delta \exp[a\theta_i], \tag{3.13}$$

and

$$\gamma^2 - 4\omega\psi = 16(a\delta)^2\big\{(\cosh^2[a\delta] - 1)\exp[2a(\theta_0 + \theta_i)]\big\}$$
$$- 16(a\delta)\big\{\sinh(2a\delta)\exp[2a(\theta_0 + \theta_i)] - \sinh[a\delta]\big(\exp[a(3\theta_0 + \theta_i)] + \exp[a(\theta_0 + 3\theta_i)]\big)\big\}. \quad (3.14)$$

Equations (3.11) – (3.14) do not appear to reduce to manageable form. However, if $\delta \to 0$, then

$$\lim_{\delta \to 0^+} \hat{b} = \frac{\theta_0 + \theta}{2}, \quad (3.15)$$

as shown in the last few lines of Appendix A.3. Therefore, items yielding optimal, expected log-likelihood ratios (for small δ and $c = 0$) have difficulty parameters, b, midway between true ability, θ_i, and the classification bound, θ_0. Figure 3.2 shows the effect of varying the c and δ on the optimal difficulty parameter. The left panels of Figure 3.2 display the optimal difficulty parameter as a function of $c \in (0, 1)$ (with δ fixed to .01), whereas the right panels indicate the optimal difficulty parameter as a function of $\delta \in (0, 1)$ (with c fixed to 0). The upper panels of Figure 3.2 fix $\theta_i = -1.0 < \theta_0 = 0$, and the lower panels fix $\theta_i = 1.0 > \theta_0$. Therefore, the optimal item difficulty parameter minimizes the log-likelihood ratio for the upper two panels of Figure 3.2 and maximizes the log-likelihood ratio for the lower two panels.

First, consider the right panels of Figure 3.2. If $c = 0$, then the b-parameter that optimizes the expected log-likelihood ratio (either minimizing if $\theta_i < \theta_0$ or maximizing if $\theta_i > \theta_0$) is close to $\frac{\theta_0 + \theta_i}{2}$, as presented in Equation (3.15). For $\delta \approx 1$, then the optimal b-parameter tiptoes closer to θ_0 but never travels much beyond the midpoint of θ_i and θ_0. Unfortunately, altering c affects the location of optimal b to a much greater extent than altering δ. To see the effect of c on the optimal b-parameter, examine the left panels of Figure 3.2. If $\theta_i < \theta_0$, then the b-parameter that minimizes Equation (3.10)

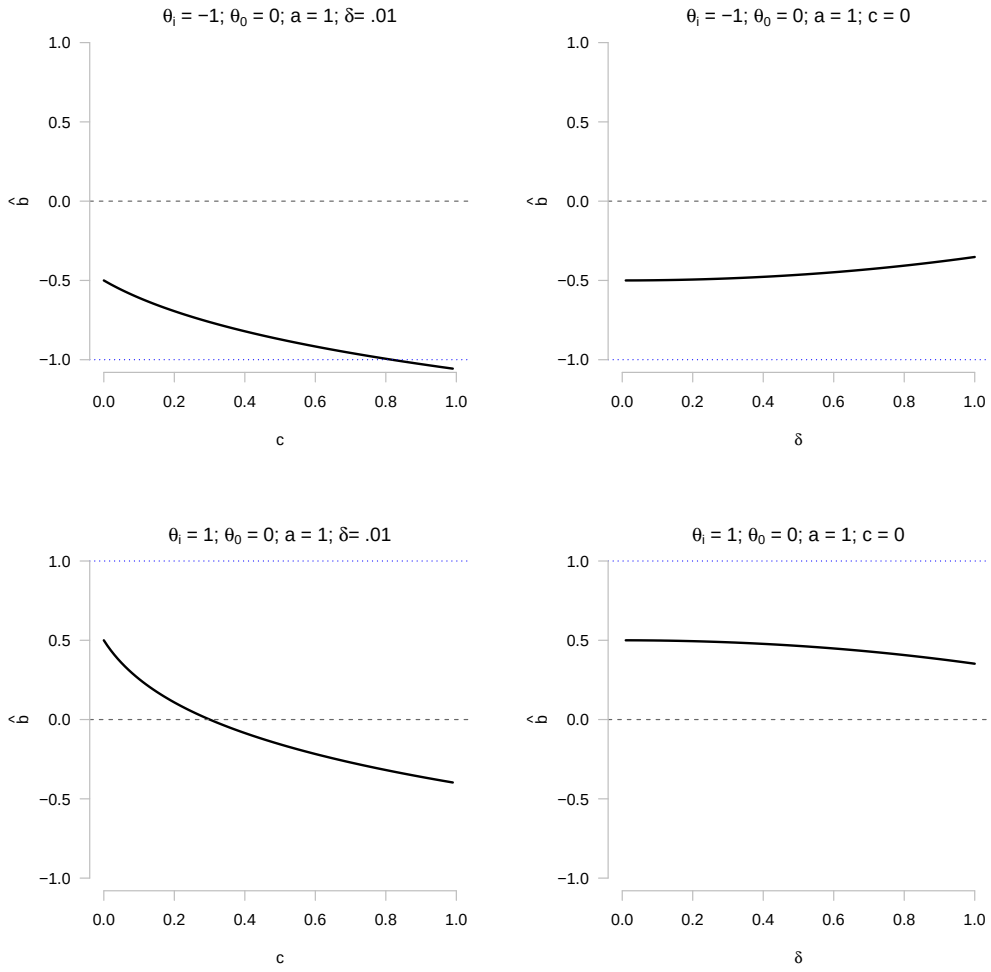

Figure 3.2: Difficulty parameters that optimize the SPRT log-likelihood ratio. The left panels show the optimal item difficulty parameter as a function of $c \in (0, 1)$, and the right panels show the optimal difficulty parameter as a function of $\delta \in (0, 1)$. The upper panels indicate the difficulty parameter that *minimizes* the expected log-likelihood ratio with respect to b for $\theta_i = -1.0 < \theta_0 = 0$ and $a = 1$, whereas the lower panels indicate the difficulty parameter that *maximizes* the expected log-likelihood ratio with respect to b for $\theta_i = 1.0 > \theta_0 = 0$ and $a = 1$.

approaches a value close to θ_i as $c \to 1$. But if $\theta_i > \theta_0$, then the b-parameter that maximizes Equation (3.10) equals (approximately) $\hat{b} = \theta_0$ for $c = .3$ and approaches a value much lower than θ_0 as $c \to 1$. Therefore, if an examinee is below the classification bound and $c > 0$, then the optimal item selection algorithm would select items closer to θ_i than θ_0, but if an examinee is above the classification bound and $c > 0$, then items should be administered with difficulty parameters close to (or even less than) θ_0.

Notice that the optimal item difficulty parameter depends on the lower asymptote (c), the size of the indifference region (δ), and the location of ability (θ_i) relative to the classification bound (θ_0). As shown in Figure 3.2, selecting items at the classification bound only approximates the ideal item selection algorithm. Under the conditions described in Figure 3.2, only for the high ability simulee with $c \approx .3$ is θ_0 the optimal item selection point. Rather than selecting items at a single point for all examinees, one could instead optimize the expected SPRT log-likelihood ratio given proximate ability estimates. Thus, in contrast to prevailing wisdom, sophisticatedly incorporating the current ability estimate into an item selection algorithm should reduce average test length relative to selecting items solely based on the classification bound.

3.1.3 The Expected SPRT Algorithm

When determining the optimal item given a fixed classification bound (or, alternatively, the optimal classification bound given a fixed item), I implicitly assumed that this item should optimize the expected SPRT criterion conditional on true ability. One could expand on these investigations by proposing an algorithm based on the expected SPRT criterion.

The expected SPRT-based item selection algorithm given a particular cut-point, θ_0, could be implemented as follows. Let $\hat{\theta}_i$ be any estimate of θ_i after $j-1$ items. For the remainder of this chapter, assume that $\hat{\theta}_i$ is found by maximum likelihood estimation.

Then an estimate of the expected SPRT-based log-likelihood ratio for the j^{th} item is

$$\begin{aligned}\text{ELR}_j(\hat{\theta}_i) &= \mathbb{E}_{\hat{\theta}_i}\left[\log\left[\text{LR}(\theta_0+\delta,\theta_0-\delta|Y_{ij})\right]\right] \\ &= p_j(\hat{\theta}_i)\log\left[\frac{p_j(\theta_0+\delta)}{p_j(\theta_0-\delta)}\right] + [1-p_j(\hat{\theta}_i)]\log\left[\frac{1-p_j(\theta_0+\delta)}{1-p_j(\theta_0-\delta)}\right], \end{aligned} \tag{3.16}$$

where $\text{ELR}_j(\hat{\theta}_i)$ stands for "the expected likelihood ratio for prospective item j given $\hat{\theta}_i$", and $p_j(\hat{\theta}_i)$ is calculated using Equation (2.1) with $\hat{\theta}_i$ inserted in place of θ_i. Finally, if $\hat{\theta}_i \geq \theta_0$, then item j should be chosen to maximize Equation (3.16), whereas if $\hat{\theta}_i < \theta_0$, then item j should be chosen to minimize Equation (3.16).

3.2 Simulation Considerations

The previous section demonstrated that the optimal difficulty parameter for a single item depends on the existence and size of the lower-asymptote in IRT models. If $c > 0$, then high ability examinees should be administered items with difficulty parameters close to θ_0, but low ability examinees should be administered items with difficulty parameters closer to θ_i than θ_0. In this section, I construct simulation studies to determine the effect of item selection, the location of the classification bound, and the magnitude of c on classification evidence for a test comprised of multiple items.

3.2.1 Simulation 1

Imagine several alternate tests with comparable item response functions but different values of c. Assume that a_1, b_1, and c_1 are known. Then to generate comparable a_2 and b_2 parameters with fixed c_2, one can minimize the squared difference between the item response functions. That is, let

$$\{a_2, b_2\} = \underset{a \geq 0, b \in \mathbb{R}}{\arg\min} \left\{ \int_{-\infty}^{\infty} \left(\Pr(Y = 1|\theta, a_1, b_1, c_1) - \Pr(Y = 1|\theta, a, b, c_2) \right)^2 \phi(\theta) d\theta \right\}, \tag{3.17}$$

where ϕ represents the standard normal density. Then $\{a_1, b_1, c_1\}$ are "similar" in item response function to $\{a_2, b_2, c_2\}$. Given this method of item construction, I generated a 1,000 item bank with $a_{1j} \sim \text{LogN}(\mu_{\log} = .53, \sigma_{\log} = .25)$, $b_{1j} \sim \text{Unif}(-4.0, 4.0)$, and $c_{1j} = .25$. I then found comparable sets of $\{a_{2j}, b_{2j}\}$ with $c_{2j} = .125$ and $\{a_{3j}, b_{3j}\}$ with $c_{3j} = 0$. Finally, classification CATs were simulated using these three item banks to classify 10,000 simulees such that $\theta_i \sim N(0, 1)$, $\delta = .1$, $\alpha = \beta = .05$, $j_{\min} = 4$, and $j_{\max} = 200$. Across the classification CATs, θ_0 was varied between -3.0 and 3.0 in 1.5 increments, and item selection was varied between maximum Fisher information at $\hat{\theta}_i$ and maximum Fisher information at θ_0.

Figure 3.3 displays the average test length for all combinations of conditions. The x-axis indicates the classification bound, and the point color type represents an item bank with a particular lower asymptote. Several aspects of Figure 3.3 are of note. First, consider the two major similarities across both plots in the upper panels. Either when $c = 0$ (the blue diamonds) or when $\theta_0 \in \{1.5, 3.0\}$ (the two right-most sets of points on either plot), the average test length remains similar regardless of whether selecting items by maximizing Fisher information at $\hat{\theta}_i$ or θ_0. As expected, the average test length only substantially differs when $c > 0$ and θ_0 is negative. The left most points on either plot clearly show the effect of $c > 0$ on average test length. When $c = .25$ and $\theta_0 = -3.0$ (e.g., trying to classify only the poorest students as in need of remedial help), the SPRT stopping rule either takes approximately 174 items (if selecting items by maximizing Fisher information at $\hat{\theta}_i$) or 24 items (if selecting items by maximizing Fisher information at θ_0) to make a decision. Note that "an average test length of 174

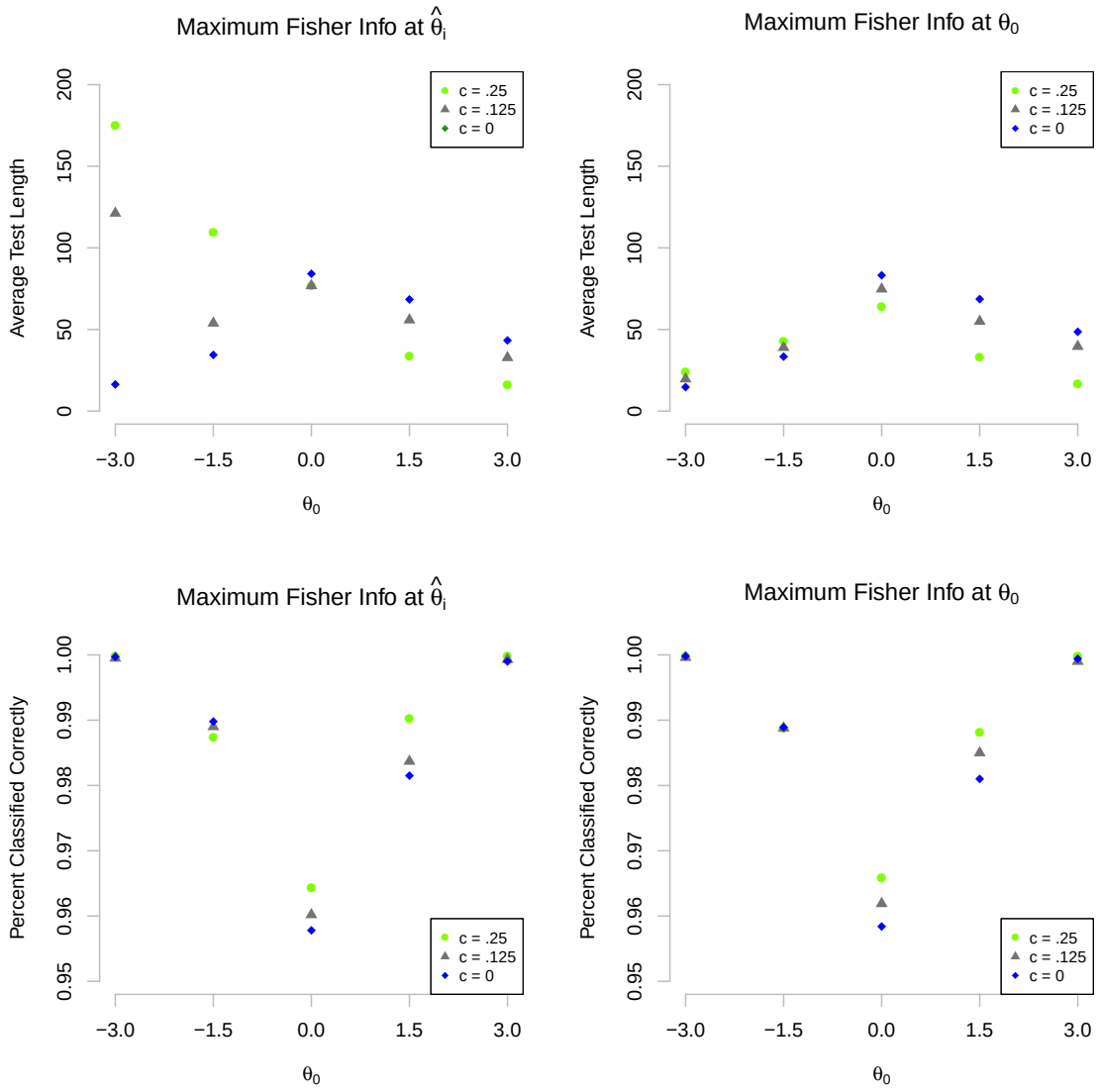

Figure 3.3: Average test length and classification accuracy using an SPRT stopping rule with different item selection algorithms, classification bounds, and lower asymptotes. For each combination of item selection algorithm, classification bound, and lower asymptote, test length was averaged across $N = 10,000$ simulees generated from a $N(0,1)$ distribution with a minimum test length of $j_{\min} = 4$ and a maximum test length of $j_{\max} = 200$. For the SPRT stopping rule, $\delta = .1$ and $\alpha = \beta = .05$.

items" severely underestimates the average number of items *required* to make an SPRT-based decision, as most of the moderate-to-high ability simulees bump up against the cap of $j_{\max} = 200$ items. Even when $c = .125$, changing the item selection algorithm to maximum Fisher information at θ_0 from maximum Fisher information at $\hat{\theta}_i$ corresponds to a dramatic decrease in the average number of items required to classify simulees above the bottom two cut-points.

Surprisingly, the upper panels of Figure 3.3 also reveal conditions in which selecting items by maximizing information at the current ability estimate is more efficient, on average, than selecting items by maximizing information at the cut-point. In fact, if $\theta_0 = 3.0$, selecting items at $\hat{\theta}_i$ results in a shorter average test length than selecting items at θ_0 for all three of the item parameter banks (16.1 versus 16.7 items, on average, for the bank with $c = .25$, 33 versus 40 items, on average, for the bank with $c = .125$, and 43 versus 49 items, on average, for the bank with $c = 0$). These results reinforce conclusions drawn from the upper left quadrant of Figure 3.2: given an examinee with low ability relative to the cut-point, one should select items closer to true ability than the cut-point. Note that if $\theta_0 = 3.0$, then practically all simulees have low ability relative to the cut-point.

The bottom two panels of Figure 3.3 display the classification accuracy rates corresponding to each of the upper plots. Notice that using either of the item selection algorithms for a given classification bound by item bank results in nearly identical classification accuracies. Therefore, the increased number of items due to selecting items by maximizing Fisher information at $\hat{\theta}_i$ does not result in a concurrent increase in classification accuracy. In fact, with few exceptions, those conditions resulting in fewer items being selected given a particular classification bound also coincide with higher classification accuracy rates.

One final aspect of Figure 3.3 merits comment. Surprisingly, the average test length

decreased for classification bounds of $\theta_0 = 1.5$ or $\theta_0 = 3.0$ when using an item bank with increased c-parameter. Although strange, these results are an artifact of item bank generation. Items were generated according to the 3PL with $c_{1j} = .25$ and corresponding item parameters (with varying lower asymptotes) were determined via Equation (3.17). The distribution used to weight the squared difference in item response functions, ϕ, up-weighted values close to 0 and down-weighted extreme values. This weighting method resulted in an adequate match of low-to-moderate difficulty items. However, items of high difficulty were weighted heavily on the lower asymptote portion of the item response function, and as a result, required much smaller item discrimination parameters to compensate. In all cases, the expected log-likelihood ratio maximally increases if items are administered with difficulty parameters between θ_i and θ_0 (review Figure 3.2). Due to the inferior set of difficult items, the banks with lower asymptotes of $c = 0$ or $c = .125$ did not have sufficiently informative items to efficiently classify moderate-to-high ability examinees below the highest classification bounds.

3.2.2 Simulation 2

I had earlier proposed an item selection algorithm based on optimizing the expected log-likelihood ratio given the current ability estimate. This item selection algorithm was motivated by derivations and graphics (using the expected log-likelihood ratio as evidence) showing how SPRT-based evidence for classification depends on both the cut-point and the true latent trait. One could also test whether an expected SPRT-based item selection algorithm results in decreased test length and equivalent classification accuracy when compared to standard CCT item selection algorithms.

Based on Chapter 2, one finds several algorithms proported to be efficient for classification CATs, including: Fisher information (FI) at θ_0, Kullback Leibler (KL) divergence between $\theta_0 + \delta$ and $\theta_0 - \delta$, and the expected log-likelihood ratio (ELR) given $\hat{\theta}_i$. Two

methods, Fisher information at θ_0 and KL divergence, only consider the classification bound, but the third method, the ELR, also requires an estimate of θ_i. To determine whether the ELR item selection method effects efficient and accurate classification tests, I simulated CCTs using the same item banks as earlier described to classify 10,000 simulees with $\theta_i \sim N(0, 1)$ and such that $\delta = .1$, $\alpha = \beta = .05$, $j_{\min} = 4$, and $j_{\max} = 200$. Across the classification CATs, θ_0 was varied between -3.0 and 3.0 in 1.5 increments, and item selection was varied between maximum Fisher information at θ_0, maximum KL divergence between $\theta_0 + \delta$ and $\theta_0 - \delta$, and optimum ELR given $\hat{\theta}_i$. Optimum ELR was defined earlier and selects items to maximize the expected log-likelihood ratio if $\hat{\theta}_i \geq \theta_0$ and minimize the expected log-likelihood ratio if $\hat{\theta}_i < \theta_0$.

Figure 3.4 displays the average test length and classification accuracy for all combinations of conditions. The x-axis indicates the classification bound, and the point color and type represents a particular item selection method. The upper panels of Figure 3.4 present results from the item bank with $c = .25$, the middle panels present results from the item bank with $c = .125$, and the lower panels present results from the item bank with $c = 0$. The left panels display the test length averaged across all examinees within each condition, and the right panels display the corresponding classification accuracies.

The patterns presented in Figure 3.4 are persistent across all three item banks. First, the expected log-likelihood ratio item selection method (the light-blue dots) always results in shorter tests than either Fisher information at θ_0 (the brown diamonds) or KL divergence (the red triangles). This relative efficiency of the ELR method increases as the classification bound shifts away from $\theta_0 = 0$. For instance, when $\theta_0 = 0$, the SPRT stopping rule takes approximately 63 (assuming $c = .25$), 73 (assuming $c = .125$), or 83 (assuming $c = 0$) items if selecting items by maximizing Fisher information at θ_0, as compared to approximately 61, 72, or 82 items, respectively, when selecting items by optimizing the ELR. However, when $\theta_0 = 3.0$, the SPRT stopping rule takes

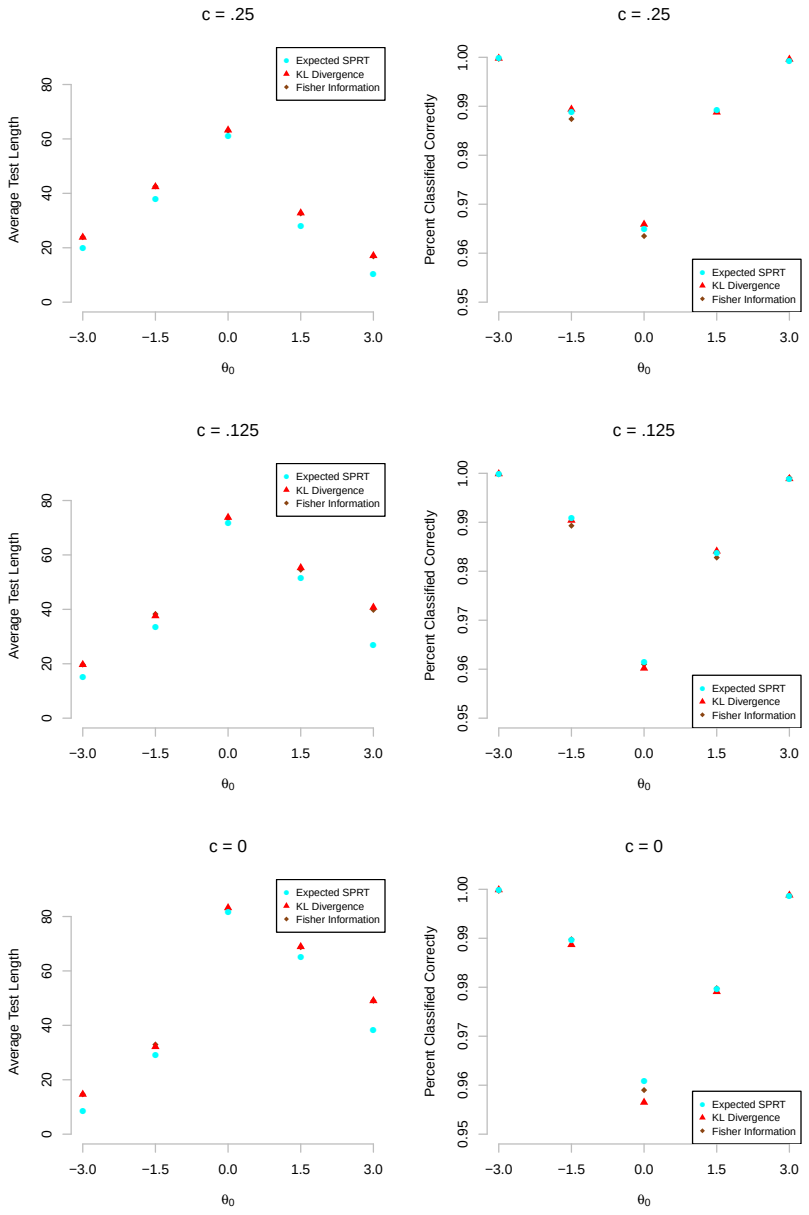

Figure 3.4: Average test length and classification accuracy using an SPRT stopping rule with different classification bound based item selection algorithms, classification bounds, and lower asymptotes. For each combination of item selection algorithm, classification bound, and lower asymptote, test length (left panels) and classification accuracy (right panels) were averaged across $N = 10,000$ simulees generated from a $N(0, 1)$ distribution with a minimum test length of $j_{\min} = 4$ and a maximum test length of $j_{\max} = 200$. For the SPRT stopping rule, $\delta = .1$ and $\alpha = \beta = .05$.

approximately 17, 40, or 49 items if selecting items by maximizing Fisher information at θ_0, as compared to approximately 10, 27, or 38 items when selecting items by optimizing the ELR. Thus, the ELR results in tests of between 1-2 items shorter for classification bounds in the center of the distribution but between 7-13 items shorter for cut-points far above the average ability. Second, the KL divergence and Fisher information at θ_0 item selection algorithms result in tests of similar length for each classification bound by item bank condition. As shown in the left panels of Figure 3.4, the red triangles generally obstruct the brown diamonds. Although both methods result in a similar number of items for each condition, maximizing KL divergence is slightly more efficient for $\theta_0 < 0$, and maximizing Fisher information at the cut-point is slightly more efficient for $\theta_0 > 0$. Note that if $\theta_0 < 0$, then most simulees are in the upper category, and the KL divergence index, as defined in Equation (2.24), more accurately reflects the true location of the average simulee. Finally, all three item selection methods result in similar accuracy rates for each classification bound by item bank condition, as shown in the right-hand panels of Figure 3.4. But when the methods diverge in accuracy rates (if $\theta_0 = 0$ and $c = 0$, for example), then the ELR method leads to either the most accurate or nearly the most accurate classifications.

As shown in this chapter, items yielding optimal SPRT evidence for classification depend on the location of true ability, θ_i, relative to the classification bound, θ_0. Selecting items by maximizing information at the classification bound is frequently not optimal. However, the increase in average test length by selecting items at the classification bound as compared to selecting items at the optimal location of b is much smaller than when selecting items at proximal estimates of θ_i. Yet selecting items by optimizing the expected log-likelihood ratio with respect to an ability estimate improves over alternative methods regardless of classification bound or item bank and with no loss in classification accuracy.

Although many researchers suggest selecting items to maximize information at θ_0, the results presented in this chapter are still counterintuitive. The recommendation to select items at the classification bound should arise from the desire to elicit the most informative *expected* response. If a test administrator gives an examinee a highly difficult item (well beyond the cut-point) and the examinee incorrectly responds to the item, then the test administrator learns very little about the examinee's ability relative to the cut-point. However, as clearly laid out in Equation (3.6), even if the examinee correctly responds to the very difficult item, the test administrator learns very little about the examinee's ability relative to the cut-point. One could not classify Einstein as a master in introductory physics if Einstein correctly responded to all questions of an advanced physics exam. But a better item selection algorithm would consider both the ability of an examinee as well as the classification bound.

In the next chapter, I present the compensatory, multidimensional IRT model as a generalization of the unidimensional 3PL IRT model. Because the functional form of both models are similar, one could apply item selection lessons derived from unidimensional IRT models in determining optimal item selection methods in multidimensional classification tests.

Chapter 4

Multidimensional Algorithms

In this chapter I briefly describe multidimensional IRT models and propose novel item selection algorithms and stopping rules for use in multidimensional mastery testing. Because one finds several generalizations of unidimensional models, I first survey diverse methods of measuring and classifying examinees among multiple dimensions before choosing a particular model for the purpose of simulations.

4.1 Multidimensional IRT and Mastery Testing

Multidimensional IRT models require added examinees for model calibration and increased computer resources for prospective adaptive testing algorithms than typical unidimensional IRT models. However, modeling item responses by incorporating additional dimensions has been shown to increase the measurement efficiency and accuracy of adaptive tests (e.g., Frey & Seitz, 2009). For instance, Segall (1996) compared a nine-dimensional, simple structure, multidimensional model against separate, unidimensional models and found that the multidimensional adaptive test required fewer items than the unidimensional adaptive tests to attain a specific SEM on each dimension. As Segall

(1996) wrote, "the gains in efficiency obtained by [multidimensional adaptive tests] depend on the correlations among the dimensions ... the larger the magnitude of these correlations, the higher the gains in efficiency over [unidimensional adaptive tests]" (p. 347). These results have been replicated across a variety of conditions (e.g., Wang & Chen, 2004, as cited in Frey & Seitz, 2009) or measurement models (e.g., Segall, 2001, who added a second-order factor to capture the correlation between the lower-order dimensions). Multidimensional models (and adaptive tests) yield better estimates of multidimensional traits than separate unidimensional scales because "when the dimensions measured by a test or battery are correlated, responses to items measuring one dimension provide clues about the examinee's standing along other dimensions" (Segall, 2000, p. 53).

Although multidimensional IRT models are increasingly recommended for use in precision-based adaptive tests, few researchers have applied mastery testing algorithms to multidimensional IRT models. The first attempt at multidimensional mastery testing (Spray, Abdel-Fatah, Huang, & Lau, 1997) approximated a multidimensional item bank with a set of unidimensional item parameters and a unidimensional function separating masters from non-masters. Despite finding sufficiently high classification accuracy across all conditions (.93–.98), Spray et al. (1997) did not compare their unidimensional approximation with the appropriate multidimensional algorithm. In contrast to Spray et al. (1997), Glas and Vos (2010) outlined a basic procedure for multidimensional mastery testing (MCMT) and found that their multidimensional algorithm resulted in increased efficiency relative to unidimensional approximations. Unfortunately, results from Glas and Vos (2010) are tempered by model choice and stopping rule. Glas and Vos (2010) built their MCMT algorithm around a multidimensional version of the one-parameter logistic model (see Equation 2.3). Moreover, they chose to use Bayesian decision theory, which requires specification of a (fairly arbitrary) loss function. The

most recent study of multidimensional mastery testing was undertaken by Seitz and Frey (2013). Unlike the aforementioned studies, Seitz and Frey (2013) applied the SPRT to multidimensional adaptive testing and compared their results to corresponding unidimensional algorithms. Unsurprisingly, accounting for multiple dimensions resulted in slightly shorter tests and more accurate classifications than ignoring the relationship between dimensions. Yet Seitz and Frey (2013) only considered one (mastery on all dimensions) version of the classification problem, one (the SPRT) stopping rule, and an inefficient (maximize the determinant of Fisher information at the current ability estimate) item selection algorithm. In the following sections, I describe a more general conceptualization of multidimensional mastery and propose novel item selection algorithms and stopping rules designed to better consider multidimensional space.

One can also classify examinees on multiple dimensions by appropriating diagnostic classification models. Unlike multidimensional item response theory (MIRT) models, diagnostic classification models (DCM) assume that the latent space consists of dichotomous or polytomous skills that combine to form K-dimensional, discrete cognitive states. Recently, DCMs have been proposed as an alternative to MIRT models for use in adaptive testing algorithms (e.g., Cheng, 2009; Gierl & Zhou, 2008; McGlohen & Chang, 2008). Unlike MCMT, which must assume that "passing" a test requires examinees to be in a particular region of multidimensional space, DCMs quantify "passing" as to whether or not an examinee evidences a certain constellation of requisite attributes. In the following sections of this chapter, I discuss various multidimensional item response theory models and contrast continuous trait conceptions of mastery with those based on discrete states. I then outline novel stopping rules and item selection algorithms for use in multidimensional mastery tests.

4.1.1 Multidimensional Item Response Theory Models

The most common generalization of the unidimensional binary response model (Equation 2.1) to multiple ability dimensions assumes that the log-odds of response is a linear function of latent ability. For instance, let $\boldsymbol{\theta}$ be the multidimensional latent variable underlying responses to test items, assume that responses are conditionally independent (see Footnote 1) given a fixed $\boldsymbol{\theta} = \boldsymbol{\theta}_i$, and allow all responses to be scored either 0 or 1. Then the probability of examinee i correctly responding to item j is often defined by the following IRF:

$$p_j(\boldsymbol{\theta}_i) = \Pr(Y_{ij} = 1|\boldsymbol{\theta}_i, \mathbf{a}_j, d_j, c_j) = c_j + \frac{1 - c_j}{1 + \exp[-(\mathbf{a}_j^T\boldsymbol{\theta}_i + d_j)]}, \tag{4.1}$$

where $\mathbf{a}_j$ represents the multidimensional slope, and all item parameters are analogous to those defined in Equation (2.1). Equation (4.1) is typically referred to as a compensatory multidimensional IRT (C-MIRT) model. Reckase (1985; also Ackerman, 1994) described how one can find multidimensional correlates of discrimination ($\text{MDISC}_j = \sqrt{\mathbf{a}^T\mathbf{a}}$), the signed distance from the origin corresponding to the line of maximum slope ($T_j = \frac{-d_j}{\text{MDISC}_j}$), and the angle with respect to arbitrary axis θ_k coinciding with the maximum slope on that axis ($\alpha_j = \arccos\left(\frac{a_{jk}}{\text{MDISC}_j}\right)$)[1].

Hooker, Finkelman, and Schwartzman (2009) critiqued the use of compensatory multidimensional IRT models. Specifically, they noted that "in the popular class of linearly compensatory models, *every* nonseparable test has a response sequence for which maximum likelihood estimates of abilities are paradoxical" (p. 420). By paradoxical, they simply meant that when using a C-MIRT model where items can load on more than one dimension, "the estimate of ability [on one dimension] can either be made to increase by changing a correctly answered item to incorrect, or to decrease by changing

[1] Although Reckase (1985) derived these relationships for models without lower asymptotes, the properties also hold if $c_j > 0$.

an incorrectly answered item to correct" (p. 20, italics in original). Because students generally assume that a correct answer should result in an increased exam score, these test properties seem paradoxical and difficult to justify. van der Linden (2012) showed that as long as Equation (4.1) defines the fundamental form of the IRT model, then compensation among ability dimensions necessarily follows. Common alternatives to C-MIRT models that do not suffer from paradoxical properties include the partially compensatory class of IRT models.

Equation (4.1) assumes that latent ability combines in a linear or compensatory fashion to predict item responses. These compensatory (or disjunctive) models predict that if an item loads on K dimensions, then high ability on one of those dimensions compensates for lower abilities on the other $K-1$ dimensions. Alternatively, one could assume that sufficient levels of all traits underlying responses to an item are required for a high probability of correctly responding to that item. These partially compensatory (or conjunctive) models assume that a correct answer to an item evinces mastery on all of the attributes comprising the item. Assume a multidimensional trait vector, $\boldsymbol{\theta}$, and conditionally independent 0–1 responses given fixed $\boldsymbol{\theta} = \boldsymbol{\theta}_i$. Then the partially compensatory multidimensional IRT model (PC-MIRT; Bolt & Lall, 2003) defines the probability of examinee i correctly responding to item j as

$$p_j(\boldsymbol{\theta}_i) = \Pr(Y_{ij} = 1|\boldsymbol{\theta}_i, \mathbf{a}_j, \mathbf{b}_j, c_j) = c_j + (1 - c_j) \prod_{k=1}^{K} \left(\frac{1}{1 + \exp[-a_{jk}(\theta_{ik} - b_{jk})]} \right)^{q_{jk}}, \tag{4.2}$$

where $q_{jk} = 1$ if item j loads on dimension k (and $q_{jk} = 0$ otherwise), a_{jk} and b_{jk} represent the k^{th} discrimination and difficulty parameter of the j^{th} item, θ_{ik} is the k^{th} element of $\boldsymbol{\theta}_i$, and $k = 1, 2, \ldots, K$ indexes dimension. Referring to $\frac{1}{1+\exp[-a_{jk}(\theta_{ik}-b_{jk})]}$ as the k^{th} component probability of the j^{th} item (see Embretson, 1984, p. 178), then

the PC-MIRT model assumes that unless every item-specific component probability is sufficiently large, then an examinee would have small probability of correctly responding to the item. Both Bolt and Lall (2003) and Babcock (2011) were able to estimate parameters for the PC-MIRT model, but Babcock (2011) commented that "the [partially compensatory] model requires a large [sample size]" (p. 327) and "the model functioned best when the latent traits had a low true [ability] correlation" (p. 327). Regardless of model type, IRT typically assumes a continuous (or semi-continuous) latent trait. Probability increases in Equation (4.1) or (4.2) if examinee i has *more* of an attribute along a particular dimension. In the next sub-section, I describe analogous response models that conceptualize the latent ability underlying responses to items as a discrete collection of (on/off, yes/no, have/have not) skills.

4.1.2 Multidimensional Diagnostic Classification Models

Typical multidimensional IRT models quantify ability as composed from continuous latent traits. One could alternatively conceptualize a multidimensional attribute vector as comprising a constellation of 0–1 discrete states. Diagnostic classification models (DCM; Rupp & Templin, 2008) quantify the latent space as a series of dichotomous (or polytomous) attributes. Rupp and Templin (2008) and Rupp, Templin, and Henson (2010) overviewed the most common DCMs, the use of DCMs in attribute testing, and the most common methods of estimating parameters of different DCMs. They also compared diagnostic classification with other common latent variable models, such as the IRT models described in Equations (4.1) and (4.2). Unlike IRT models, which generally possess little within-item multidimensionality, DCMs contain "latent variables that typically operationalize more narrowly defined constructs – so that each item requires multiple component skills" (Rupp & Templin, 2008, p. 230). Because of the difference in overall and within-item dimensionalities, DCMs are known to poorly retrofit

assessments originally designed for broadly-based, IRT traits (Gierl & Cui, 2008).

As in IRT, DCMs are typically divided into conjunctive and disjunctive models. Conjunctive models require that an examinee possesses all attributes comprising an item to have a high probability of correctly responding to that item. Common conjunctive models include the deterministic input noisy-and-gate (DINA; Junker & Sijtsma, 2001), the noisy input determinstic-and-gate (NIDA; Junker & Sijtsma, 2001), and the fusion model (Rousseau et al., 2007). The DINA models item response probabilities with item-specific slipping (i.e., the probability of an incorrect response given an examinee with all of the required attributes for solving an item) and guessing parameters, whereas the NIDA models these probabilities with attribute-specific slipping and guessing parameters. Neither the DINA nor NIDA model posits an interaction between attributes and item difficulty save for defining the attributes required for solving an item. The fusion model includes both an interaction between attributes and items and an additional parameter that accounts for auxiliary attributes. Specifically, let α_{ik} indicate whether examinee i has attribute k, let s_{jk} designate the probability of an examinee with attribute k incorrectly responding to part of an item requiring attribute k, let g_{jk} represent the probability of an examinee without attribute k correctly responding to the part of an item requiring attribute k, and let $q_{jk} = 1$ if item j requires attribute k. Then the fusion model defines two additional parameters, π_j and r_{jk}, such that

$$\pi_j = \prod_{k=1}^{K} \Pr(Y_{ijk} = 1 | \alpha_{ik} = 1) = \prod_{k=1}^{K} (1 - s_{jk})^{q}_{jk} \tag{4.3}$$

and

$$r_{jk} = \frac{\Pr(Y_{ijk} = 1 | \alpha_{ik} = 0)}{\Pr(Y_{ijk} = 1 | \alpha_{ik} = 1)} = \frac{g_{jk}}{1 - s_{jk}}. \tag{4.4}$$

Equation (4.3) defines the probability of an examinee correctly responding to an item

given all of the attributes required for that item, and Equation (4.4) defines the penalty accrued (relative to the perfect examinee) by an examinee not having attribute k. Using Equation (4.3) and (4.4), the fusion model is defined as

$$p_j(\theta_i, \boldsymbol{\alpha}_i) = \Pr(Y_{ijk} = 1|\theta_i, \boldsymbol{\alpha}_i, b_j, \pi_j, \mathbf{r}_j) = \frac{\pi_j \prod_k r_{jk}^{(1-\alpha_{ik})q_{jk}}}{1 + \exp(\theta_i - b_j)}, \tag{4.5}$$

where θ represents a continuous latent trait that limits the probability of response for a particular examine, and all other terms were defined above. The additional parameters, π_j and r_{jk}, were constructed from s_{jk} and g_{jk} due to identifiability concerns (e.g., Wang, Chang, & Huebner, 2011, p. 257).

Unlike conjunctive models, disjunctive DCMs require examinees to only possess one of the attributes composing an item to have a high probability of correctly responding to that item. Common disjunctive DCMs include the deterministic noisy-or-gate (DINO), noisy input deterministic-or-gate (NIDO), and the compensatory, reparameterized unified model (C-RUM) (see Rupp & Templin, 2008). Not surprisingly, DINO defines the probability of response by modeling item characteristics, whereas NIDO defines the probability of response be modeling attribute characteristics. Moreover, many disjunctive DCMs are discrete analogues of standard, multidimensional IRT models. For instance, let ζ_{i0} be the log-odds of an examinee answering item i correctly without any of the required attributes, let ζ_{ik} be the gain in log-odds of an examinee answering an item correctly if he/she has attribute k, and define all other terms as in Equation (4.5). Then the C-RUM defines the probability of a correct response for examinee i to item j as

$$p_j(\boldsymbol{\alpha}_i) = \Pr(Y_{ij} = 1|\boldsymbol{\alpha}_i, \boldsymbol{\zeta}_j) = \frac{1}{1 + \exp[-(\zeta_{j0} + \sum_k \zeta_{jk}\alpha_{ik}q_{jk})]}. \tag{4.6}$$

Equation (4.6) is very similar to Equation (4.1) with the continuous latent trait vector

replaced by discrete attributes.

DCMs have been proposed as alternate measurement models for adaptive tests. Therefore, one could develop mastery testing algorithms corresponding to DCMs. In contrast to IRT, DCMs would directly quantify the probability of mastery (assuming mastery is defined as a set of multidimensional attribute vectors) and, thus, require little modification for use in mastery testing algorithms (although, see Rupp & Templin, 2008, p. 235, for an argument against using DCMs for classification). Thus, henceforth, I focus on developing and testing multidimensional classification algorithms using the IRT measurement model with continuous latent trait vectors. Any mention of DCMs will only be for comparative purposes. In the next sub-section, I briefly summarize multidimensional conceptions of mastery using IRT models.

4.1.3 Multidimensional Mastery Testing

Multidimensional computerized mastery testing (MCMT) requires algorithms to determine when an examinee's latent trait is located within a pre-specified region of multidimensional space. These regional definitions can also be used to determine the optimal item selection rules for differentiating two examinees slightly within each region. For example, Chapter 3 shows that items should be selected primarily based on the cut-point separating categories. Because the multidimensional compensatory IRT model, as defined in Equation (4.1), is similar in form to the 3PL model, one should also pick multidimensional mastery items based on the boundary between mastery and non-mastery.

Very little work has extended mastery testing to multidimensional problems. The first paper to discuss multidimensional mastery testing, Spray et al. (1997), quantified mastery based on a minimally competent percentage of correct responses, $p_0 = \sum_j p_j(\boldsymbol{\theta}_0)$. If p_0 is determined beforehand and the sample sequence of items is known,

then $\boldsymbol{\theta}_0$ divides the latent space into two regions: a mastery region, in which the percentage of correct responses is typically greater than p_0, and a non-mastery region, in which the percentage of correct responses is typically less than p_0. As Spray et al. (1997) noted, the values of $\boldsymbol{\theta}_0$ that satisfy $p_0 = \sum_j p_j(\boldsymbol{\theta}_0)$ define a curve in $\mathbb{R}^K$, where K is the dimension of $\boldsymbol{\theta}$. To illustrate the passing region described by Spray et al. (1997), I generated $J = 40$ parameters to fit a two-dimensional C-MIRT model with $\bar{a}_1 = .81$ ($s_{a_1} = .59$), $\bar{a}_2 = .84$ ($s_{a_2} = .61$), $\bar{d} = -.53$ ($s_d = .82$), and $c = 0$. I then determined (θ_1, θ_2) pairs that would result in average, model-predicted probabilities of $p_0 = .4$, $p_0 = .6$, and $p_0 = .8$. The resulting classification bound functions are plotted in Figure 4.1. Note that for $p_0 = .4$, and $p_0 = .8$, the threshold functions define non-linear curves in two-dimensional space.

As originally proposed by Spray et al. (1997), constructing these constant probability classification bounds requires an unchanging set of parameters and a fixed model. Different models will yield different mastery regions. One could, of course, define the mastery region based on a test set of items and then interpolate a curve between those points to use with alternative item banks or IRT models. However, Glas and Vos (2010) argued that the passing region should not necessarily be directly related to the underlying model. According to Glas and Vos (2010), "the choice of compensatory or non-compensatory model is an empirical matter, ... [whereas] the choice of ... [classification region] is a value judgment determined by the opinion of who can be qualified as a master" (p. 429). In other words, responses to mathematical comprehension items might (empirically) be determined by a linear combination of reading and computational abilities, but examinees might still need sufficient ability on both dimensions to qualify as a master. Disconnecting the mastery decision from the item response function, Glas and Vos (2010) defined two types of classification procedures. A non-compensatory (or conjunctive) classification procedure requires examinees to be above a threshold on all

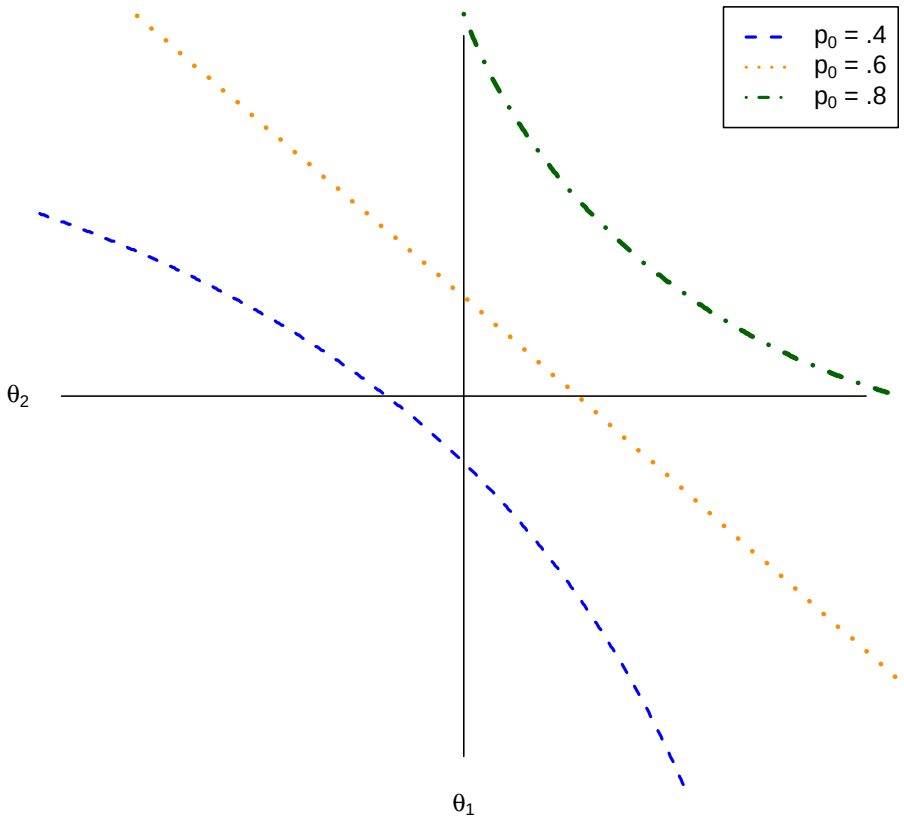

Figure 4.1: Classification bound functions assuming a minimal, constant, model-predicted probability for passing the test. Probabilities were generated using the two-dimensional C-MIRT model with $\bar{a}_1 = .81$, $\bar{a}_2 = .84$, $\bar{d} = -.53$, and $c = 0$.

dimensions to qualify as a master. An example of a two-dimensional, non-compensatory classification task is provided in Figure 4.2. One could modify non-compensatory classification regions for use in diagnostic classification modeling by requiring the posterior probability of an examinee on each of the required attributes to exceed some threshold. Conversely, a compensatory classification procedure requires a linear combination of an examinee's traits to be above a threshold for the examinee to qualify as a master. An examinee of a two-dimensional, compensatory classification task is provided in Figure 4.3.

Glas and Vos (2010) proposed compensatory and non-compensatory classification regions for constructing loss functions in multidimensional space. Once loss functions were defined, they used Bayesian decision theory to both select items and make classification decisions and found that multidimensional CMT improved over a unidimensional analogue as the correlation between the dimensions decreased.

The most recent conception of multidimensional CMT was described by Seitz and Frey (2013). As in Spray et al. (2011), Seitz and Frey (2013) were unable to generalize the SPRT stopping rule without severe restrictions on the item bank and the classification function. For instance, Seitz and Frey (2011) chose an item bank with between-item unidimensionality. Because they assumed that each item only loaded on one dimension, they simplified the classification task by comparing every $\theta_{0k} + \delta$ against $\theta_{0k} - \delta$, where θ_{0k} is the cut-point for dimension k. Therefore, the SPRT described by Seitz and Frey (2013) contrasts the specific hypotheses: $\mathrm{H}_0 : \boldsymbol{\theta}_i = \boldsymbol{\theta}_0 - \delta$ and $\mathrm{H}_1 : \boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \delta$. Because the point hypotheses are the same for all examinees and all items, these authors avoid constructing a mastery region or considering the distance between each examinee's trait level and the border of that region. Moreover, as shown in Figures 4.2 and 4.3, testing $\boldsymbol{\theta}_0 + \delta$ against $\boldsymbol{\theta}_0 - \delta$ would be consistent with non-compensatory, compensatory, or a variety of other classification bound functions. Additionally, Seitz and Frey (2013)

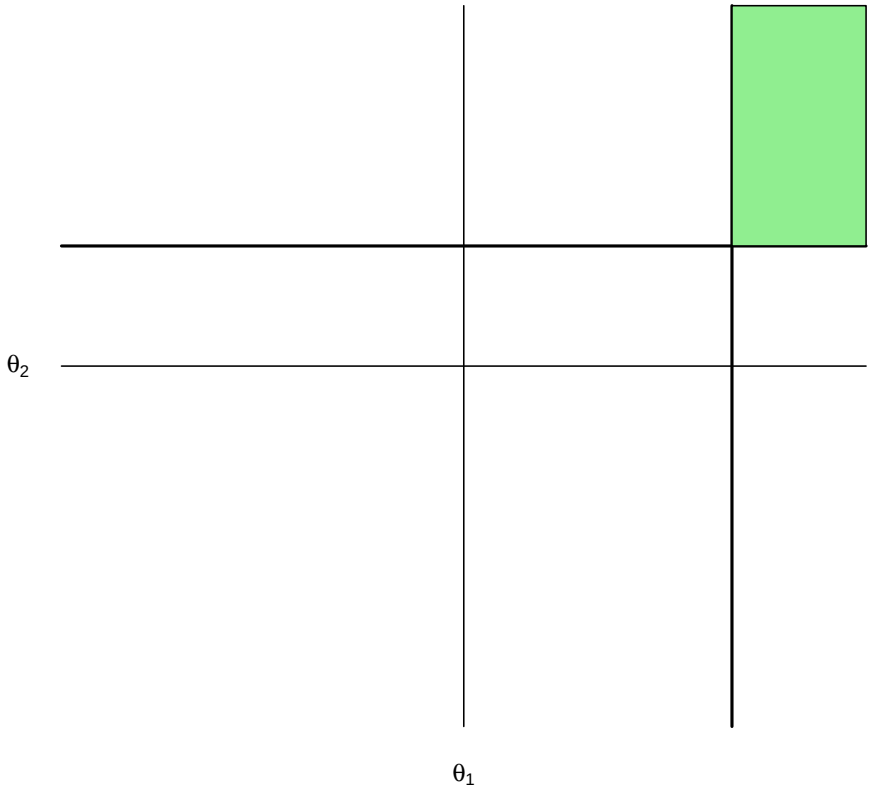

Figure 4.2: A diagram of a non-compensatory classification task. An examinee is required to be in the shaded, green box (upper-right) to be considered a master and, therefore, must be above the threshold on both dimensions.

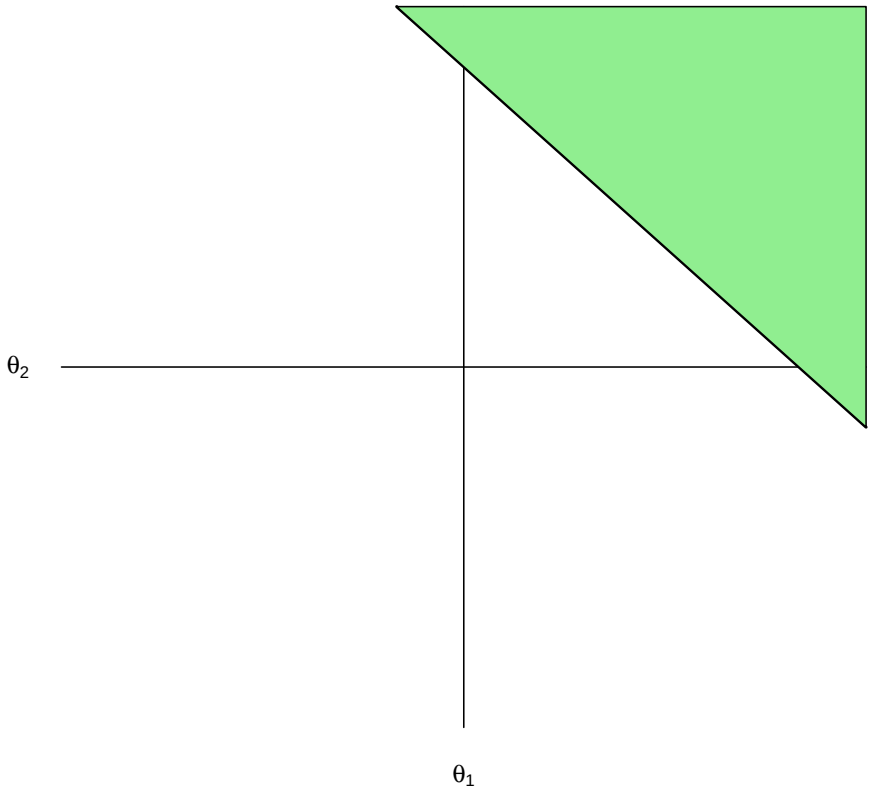

Figure 4.3: A diagram of a compensatory classification task. An examinee is required to be in the shaded, green box (upper-right) to be considered a master. However, for this task, a sufficiently high ability on one dimension would compensate for a low ability on the other dimension.

avoided developing or using item selection algorithms appropriate for mastery tests and, instead, used an algorithm synonymous with maximizing Fisher information at the current ability estimate. Thus, in the remaining sections of this chapter, I propose modified SPRT-based stopping rules and item selection algorithms appropriate for determining whether examinees are within regions of multidimensional space.

4.2 Multidimensional Stopping Rules

In the following sub-sections, I propose generalizations of unidimensional CMT stopping rules to mastery tests comprised of multiple dimensions. The compensatory MIRT model, as defined in Equation (4.1), will be used to illustrate application of the methods. However, any of the stopping rules can be used, in principle, with any MIRT model.

4.2.1 Multidimensional Sequential Probability Ratio Tests

Spray et al. (1997) noted a problem in generalizing the SPRT to multidimensional IRT models. Rather than defining point hypotheses to represent each category, one "would need to define the likelihood ratio as before along two distinct curves approximately parallel to the [classification bound]" (p. 5). However, as is immediately obvious if superimposing classification bound functions onto a contour plot of the log-likelihood function, in almost no circumstance is the likelihood ratio constant along two curves parallel to the classification bound function. Therefore, the SPRT point hypotheses must depend on the location of $\hat{\boldsymbol{\theta}}_i$. Because many pairs of points could be selected that compare values in the mastery region to values in the non-mastery region, any reasonable SPRT generalization must satisfy several criteria. Let H_1 be composed of a curve with all points δ away from the classification bound function and in the mastery region, and let H_0 be composed of a curve with all points δ away from the classification bound

function and in the non-mastery region. Then the points chosen along H_1 and H_0 to be compared in a likelihood ratio must: (1) be likely relative to other points along the curves defining the hypotheses; (2) be close to each other (in terms of likelihood) relative to other pairs of points along the curves defining the hypotheses; and (3) be close to each other (in terms of distance) relative to other pairs of points defining the curves along the hypotheses. Given these criteria, I developed two possible generalizations of the fixed-point SPRT to multiple dimensions: the Constrained SPRT (C-SPRT) and the Projected SPRT (P-SPRT).

The Constrained SPRT (C-SPRT) determines the fixed points used in the likelihood ratio test statistic by finding the maximum likelihood estimate constrained to lie on the classification bound function. Specifically, define a classification bound function, $g(\boldsymbol{\theta})$, satisfying the equality constraint $g(\boldsymbol{\theta}) = 0$. $g(\boldsymbol{\theta})$ can be a linear function, a curvilinear function, or a piece-wise function. For instance, the function

$$g(\boldsymbol{\theta}) = \theta_2 + 1.5\theta_1 - .5$$

would define the compensatory classification bound $\theta_2 = -1.5\theta_1 + .5$, whereas the function

$$g(\boldsymbol{\theta}) = \begin{cases} \theta_1 - 2 & \text{if } \theta_2 \geq 1 \\ \theta_2 - 1 & \text{if } \theta_1 \geq 2 \\ 1 & \text{otherwise} \end{cases}$$

would define the non-compensatory classification bound such that $\theta_1 > 2$ and $\theta_2 > 1$ designates masters. After item j_{tmp}, the C-SPRT algorithm would find the maximum likelihood estimate constrained to lie on the classification bound function,

$$\hat{\boldsymbol{\theta}}_0 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0}{\arg\max} \left[\log[L(\boldsymbol{\theta}|\mathbf{y}_{i,\,j_{\text{tmp}}})] \right], \tag{4.7}$$

where $\boldsymbol{\Theta}_0 := \{\boldsymbol{\theta} : \, g(\boldsymbol{\theta}) = 0\}$ and

$$\log[L(\boldsymbol{\theta}|\mathbf{y}_{i,\,J})] = \sum_{j=1}^{J} \Big[y_{ij} \log[p_j(\boldsymbol{\theta})] + (1 - y_{ij}) \log[1 - p_j(\boldsymbol{\theta})] \Big]. \tag{4.8}$$

Given $\hat{\boldsymbol{\theta}}_0$, the C-SPRT determines the line perpendicular to $g(\boldsymbol{\theta}) = 0$ and chooses values $\pm\delta$ away from $\hat{\boldsymbol{\theta}}_0$ along this line. These values are then compared in a log-likelihood ratio to pre-specified critical values.

Calculating the points to use in a likelihood ratio requires some knowledge of elementary calculus. Define $\boldsymbol{\theta}_l = \hat{\boldsymbol{\theta}}_0 - \delta\boldsymbol{\theta}_\delta$ and $\boldsymbol{\theta}_u = \hat{\boldsymbol{\theta}}_0 - \delta\boldsymbol{\theta}_\delta$ to be the lower and upper values used in a log-likelihood ratio. For these values to be appropriate, $\boldsymbol{\theta}_\delta$ should be a unit-length vector such that $\hat{\boldsymbol{\theta}}_0 + \boldsymbol{\theta}_\delta$ is on the line orthogonal to the tangent plane $\nabla g(\hat{\boldsymbol{\theta}}_0)^T[\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0] = 0$. The symbol $\nabla g(\boldsymbol{\theta})$ represents the vector of partial derivatives of $g(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. From basic calculus, $\mathbf{h}(t) = \hat{\boldsymbol{\theta}}_0 + t\nabla g(\hat{\boldsymbol{\theta}})_0$ determines the line orthogonal to the plane tangent to $g(\hat{\boldsymbol{\theta}}_0)$ (e.g., Stewart, 2007, p. 549). Therefore, $t = \frac{1}{||\nabla g(\hat{\boldsymbol{\theta}}_0)||}$, where $||\cdot||$ is the Euclidean norm, so that $\boldsymbol{\theta}_\delta = \frac{\nabla g(\hat{\boldsymbol{\theta}}_0)}{||\nabla g(\hat{\boldsymbol{\theta}}_0)||}$.

Two examples should clear up any confusion from the previous paragraph. First, assume a classification problem with $g(\boldsymbol{\theta}) = \theta_2 + 1.5\theta_1 - .5$, and let $\hat{\boldsymbol{\theta}}_0 = [1, -1]^T$. These properties define a compensatory classification problem with the line $\theta_2 = .5 - 1.5\theta_1$ separating masters from non-masters and the point $\hat{\boldsymbol{\theta}}_0 = [1, -1]^T$ on that dividing line. Then $\nabla g(\hat{\boldsymbol{\theta}}_0) = [1.5, 1]^T$, so that $\boldsymbol{\theta}_\delta = \frac{\nabla g(\hat{\boldsymbol{\theta}}_0)}{||\nabla g(\hat{\boldsymbol{\theta}}_0)||} = [1.5/\sqrt{3.25}, 1/\sqrt{3.25}]^T$. $\delta\boldsymbol{\theta}_\delta$ should be added to $\hat{\boldsymbol{\theta}}_0 = [1, -1]^T$ to determine the two points compared in a log-likelihood ratio. Second, assume a classification problem with

$$g(\boldsymbol{\theta}) = \begin{cases} \theta_2 + .1\theta_1^2 - 2 & \text{if } \theta_1 > 0 \\ \theta_1 & \text{if } \theta_2 \geq 2 \\ 0 & \text{otherwise} \end{cases},$$

and let $\hat{\boldsymbol{\theta}}_0 = [2, 1.6]^T$. These properties define a mixed compensatory/non-compensatory classification bound function so that an examinee must have true $\theta_1 > 0$ as well as having $\theta_2 > 2 - .1\theta_1^2$ for them to be considered a master. Given $\hat{\boldsymbol{\theta}}_0 = [2, 1.6]^T$ (so that $\theta_1 > 0$), the appropriate gradient would be $\nabla g(\hat{\boldsymbol{\theta}}_0) = [0.4, 1]^T$, so that $\boldsymbol{\theta}_\delta = \frac{\nabla g(\hat{\boldsymbol{\theta}}_0)}{||\nabla g(\hat{\boldsymbol{\theta}}_0)||} = [0.4/\sqrt{1.16}, 1/\sqrt{1.16}]^T$. In this case, $\delta\boldsymbol{\theta}_\delta$ should be added to $\hat{\boldsymbol{\theta}}_0 = [2, 1.6]^T$ to determine the two points compared in a log-likelihood ratio. Figure 4.4 depicts the process of finding $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_l$ to construct the C-SPRT given a particular classification bound function, MIRT model, and set of item responses. Note that the C-SPRT does not require a global estimate of $\hat{\boldsymbol{\theta}}$.

In contrast to the C-SPRT, the Projected SPRT (P-SPRT) determines the fixed points used in the likelihood ratio test statistic by projecting the unconstrained MLE orthogonally onto the closest point of the classification bound surface. As before, define a classification bound function, $g(\boldsymbol{\theta})$, satisfying the equality constraint $g(\boldsymbol{\theta}) = 0$, and let $\hat{\boldsymbol{\theta}}_{j_{\text{tmp}}}$ be the maximum likelihood estimate after j_{tmp} items. Then the projected maximum likelihood estimate after j_{tmp} items would be

$$\hat{\boldsymbol{\theta}}_0 = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0}{\arg\min} ||\hat{\boldsymbol{\theta}}_{j_{\text{tmp}}} - \boldsymbol{\theta}||, \tag{4.9}$$

where $||\cdot||$ is the Euclidean norm function. After determining $\hat{\boldsymbol{\theta}}_0$ from Equation (4.9), the P-SPRT would proceed in the same manner as the C-SPRT. Figure 4.5 depicts the process of finding $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_l$ to construct the P-SPRT given a particular classification

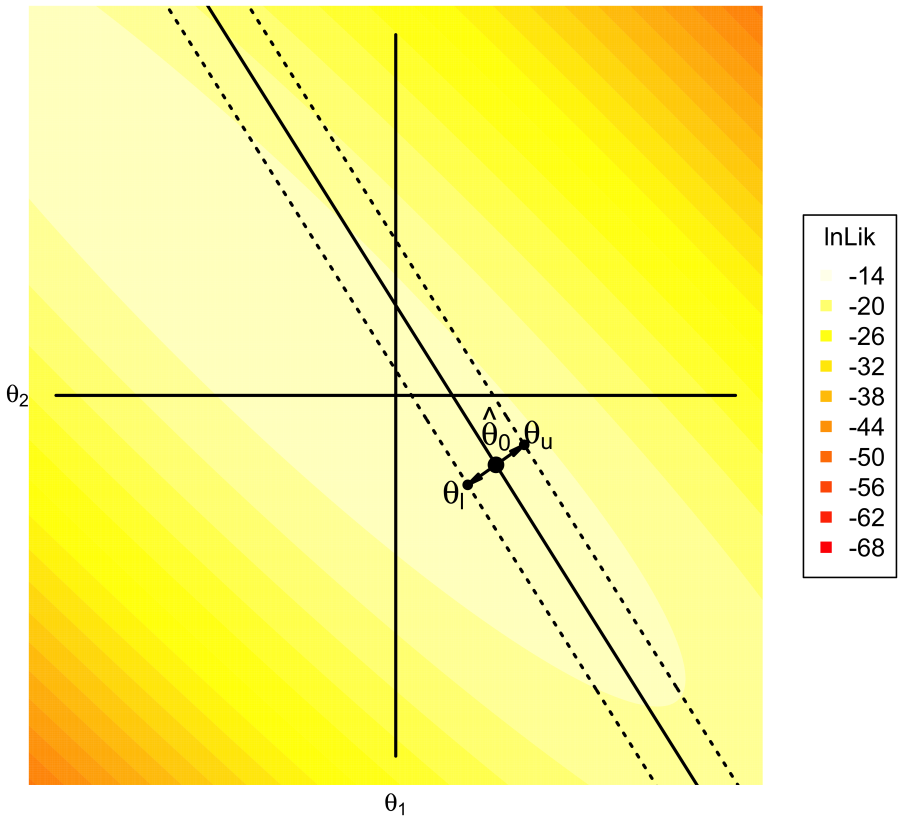

Figure 4.4: A diagram of the Constrained SPRT in two dimensions. The likelihood was constructed using the compensatory MIRT model, as defined in Equation (4.1), and the classification bound function was constructed assuming a compensatory classification task with $g(\boldsymbol{\theta}) = \theta_2 + 1.5\theta_1 - .5$.

bound function, MIRT model, and set of item responses. Note that unlike the C-SPRT, the P-SPRT does require a global estimate of $\hat{\boldsymbol{\theta}}$ to find the closest point on the classification bound function.

Once $\boldsymbol{\theta}_l = \hat{\boldsymbol{\theta}}_0 - \delta\boldsymbol{\theta}_\delta$ and $\boldsymbol{\theta}_u = \hat{\boldsymbol{\theta}}_0 - \delta\boldsymbol{\theta}_\delta$ are found for a particular classification bound function, $g(\boldsymbol{\theta})$, and a constrained ability estimate, $\hat{\boldsymbol{\theta}}_0$, the log-likelihood ratio of examinee i manifesting $\boldsymbol{\theta}_u$ relative to $\boldsymbol{\theta}_l$ would be

$$C_{i,\,j} = \log\Big[\mathrm{LR}(\boldsymbol{\theta}_u, \boldsymbol{\theta}_l|\mathbf{y}_{i,\,j}\Big] = \log\left[\frac{L(\boldsymbol{\theta}_u|\mathbf{y}_{i,\,j})}{L(\boldsymbol{\theta}_l|\mathbf{y}_{i,\,j})}\right] = \log\Big[L(\boldsymbol{\theta}_u|\mathbf{y}_{i,\,j})\Big] - \log\Big[L(\boldsymbol{\theta}_l|\mathbf{y}_{i,\,j})\Big] \quad (4.10)$$

regardless of whether using the C-SPRT or P-SPRT to determine classification.

The C-SPRT and P-SPRT pick the closest point on the classification bound in slightly different ways. The P-SPRT defines closest based on the distance between the current ability estimate and the classification bound function. In contrast, the C-SPRT defines closest based on the log-likelihood function along the classification bound. Whereas the C-SPRT can be justified using a similar rationale as to the justification underlying generalized likelihood ratios in sequential stopping problems, the P-SPRT should only result in efficient and accurate classification tests when the classification bound function roughly aligns with a contour of the log-likelihood function.

4.2.2 Multidimensional Generalized Likelihood Ratio Tests

Unlike modifications needed to generalize the SPRT to multiple dimensions (assuming, of course, hypotheses as functions rather than points), the Generalized Likelihood Ratio Test naturally generalizes to multidimensional classification problems. Define a classification bound function, $g(\boldsymbol{\theta}) = 0$, that separates a non-mastery region, $\boldsymbol{\Theta}_n$, from a mastery region, $\boldsymbol{\Theta}_m$. Then an extension of the simple GLR (e.g., Thompson, 2009, 2010) test statistic can be written as

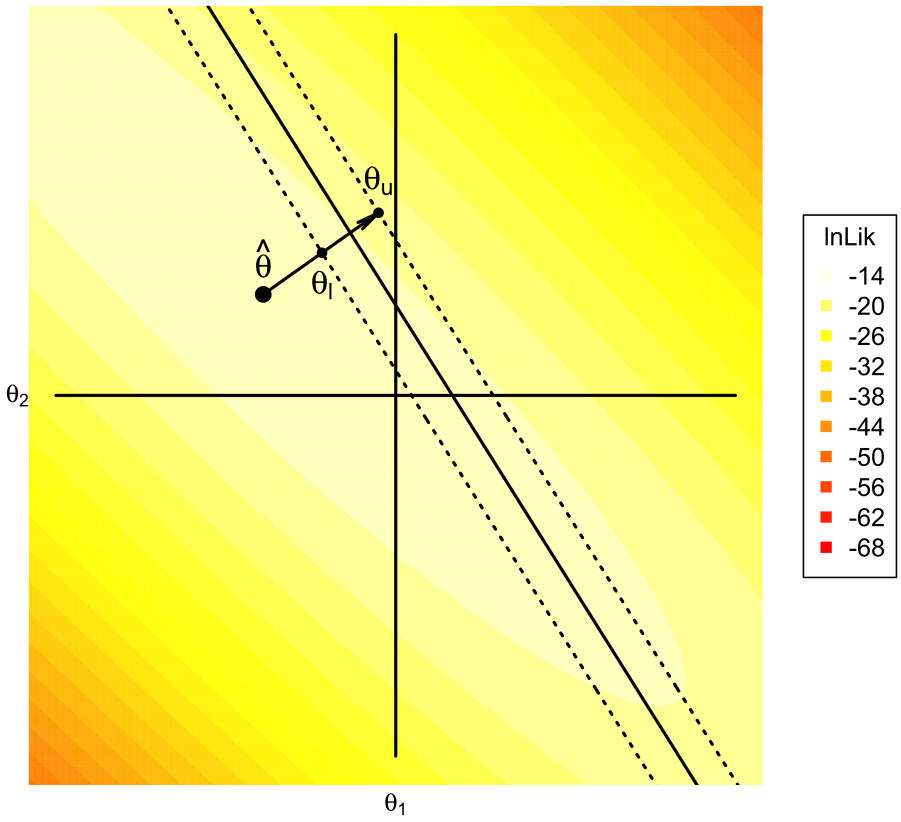

Figure 4.5: A diagram of the Projected SPRT in two dimensions. The likelihood was constructed using the compensatory MIRT model, as defined in Equation (4.1), and the classification bound function was constructed assuming a compensatory classification task with $g(\boldsymbol{\theta}) = \theta_2 + 1.5\theta_1 - .5$.

$$G_{i,j} = \sup_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_m} \left(\log \left[L(\boldsymbol{\theta}_1 | \mathbf{y}_{i,j}) \right] \right) - \sup_{\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_n} \left(\log \left[L(\boldsymbol{\theta}_2 | \mathbf{y}_{i,j}) \right] \right). \tag{4.11}$$

This Multidimensional Generalized Likelihood Ratio (M-GLR) statistic corresponds to the maximum of the likelihood conditional on being within the mastery region divided by the maximum of the likelihood function conditional on being within the non-mastery region. For simple classification bound functions, these maximums are easily found using a constrained optimization routine. Of course, as in the unidimensional case, one could define the mastery region (and non-mastery region) as restricted to lie a certain distance from the classification bound function by choosing an appropriate δ_l such that $\mathbf{h}_l(\boldsymbol{\theta}) = \boldsymbol{\theta}_0 - \delta_l \boldsymbol{\theta}_\delta$ and an appropriate δ_u such that $\mathbf{h}_u(\boldsymbol{\theta}) = \boldsymbol{\theta}_0 + \delta_u \boldsymbol{\theta}_\delta$ (where $\boldsymbol{\theta}_0$ satisfies $g(\boldsymbol{\theta}_0) = 0$). One could also find (e.g., Bartroff, Finkelman, & Lai, 2008) values to plug into the likelihood function (as well as critical values) via simulation. However, in all cases, the relevant unidimensional theory would also accommodate multidimensional classification problems. Figure 4.6 depicts points selected to construct a M-GLR given a particular classification bound function, MIRT model, and set of item responses.

Rather than comparing two fixed points, one could instead define composite hypotheses

$$\mathrm{H}_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_n$$
$$\mathrm{H}_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_m$$

and compare a weighted average of the likelihood ratio across each composite hypothesis. Weighted likelihood ratios are well-known in statistics (e.g., Dickey, 1971), and averaging the likelihood ratio across two regions relates to Bayes factors (e.g., Lachin, 1981; Lavine & Schervish, 1999). Using previous notation, let w_{ij} be a weight function for examinee

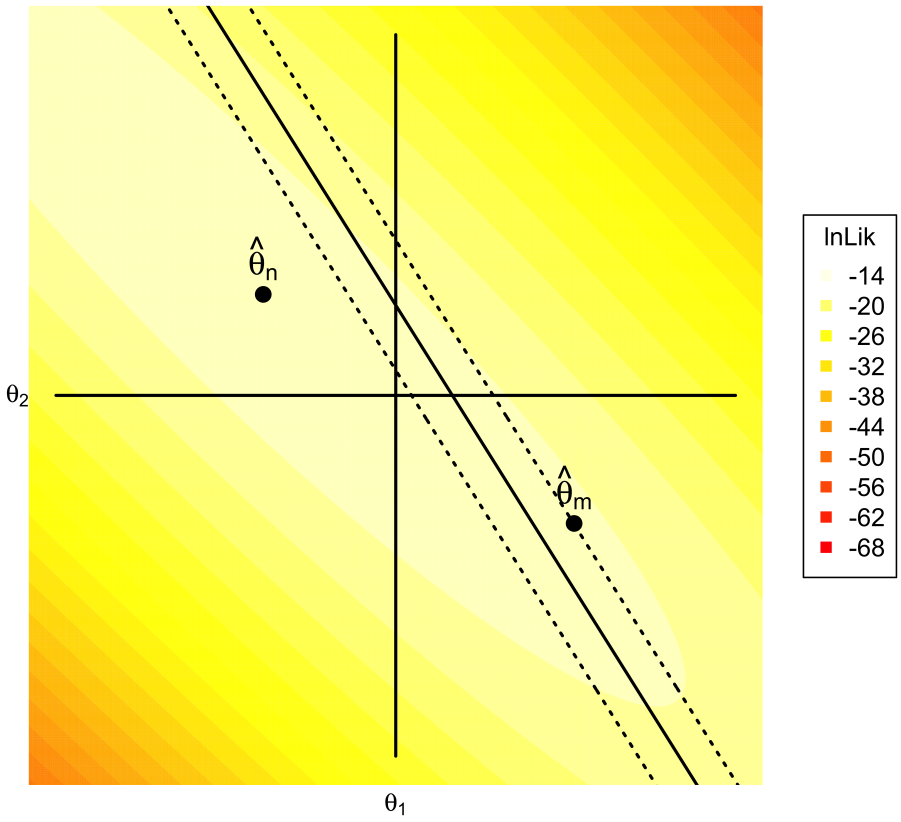

Figure 4.6: A diagram of the Multidimensional GLR in two dimensions. The likelihood was constructed using the compensatory MIRT model, as defined in Equation (4.1), and the classification bound function was constructed assuming a compensatory classification task with $g(\boldsymbol{\theta}) = \theta_2 + 1.5\theta_1 - .5$.

i after j items, and let $\mu_m(w_{ij}) = \int_{\boldsymbol{\Theta}_m} w_{ij} d\theta$ and $\mu_n(w_{ij}) = \int_{\boldsymbol{\Theta}_n} w_{ij} d\theta$. Then the Weighted GLR (W-GLR) can be defined as

$$G_{i,\,j} = \log\left[\frac{\int_{\boldsymbol{\Theta}_m} w_{ij} L(\boldsymbol{\theta}|\mathbf{u}_{i,\,j}) d\theta}{\mu_m(w_{ij})}\right] - \log\left[\frac{\int_{\boldsymbol{\Theta}_n} w_{ij} L(\boldsymbol{\theta}|\mathbf{u}_{i,\,j}) d\theta}{\mu_n(w_{ij})}\right]. \tag{4.12}$$

In practice, researchers should set w_{ij} equal to the the prior distribution of $\boldsymbol{\theta}$, in which case $G_{i,\,j}$ would be the logarithm of the Bayes factor. Jha, Clarke, Langmead, Legay, Platzer, and Zuliani (2013) proposed comparing $G_{i,\,j}$ to an a priori specified number, T, choosing hypothesis H_1 if $G_{i,\,j} > \log(T)$ and choosing hypothesis H_0 if $G_{i,\,j} < \log(1/T)$. Jha et al. (2013) found that given comparable thresholds, the resulting sequential Bayesian test needed fewer samples than the corresponding SPRT.

Berger (2012) argued that "in sequential scenarios, there is no need to 'spend α' for looks at the data" when using Bayesian tests because "posterior probabilities are not affected by the reason for stopping experimentation" (p. 49). In this case, one would simply calculate the posterior probability of being in the mastery or non-mastery regions and choose a hypothesis if the posterior probability of that hypothesis is greater than some $1 - \alpha$. For example, let

$$\pi(\boldsymbol{\theta}|\mathbf{y}_{i,\,j_{\mathrm{tmp}}}) = \frac{\pi(\boldsymbol{\theta}) L(\boldsymbol{\theta}|\mathbf{y}_{i,\,j_{\mathrm{tmp}}})}{\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}) L(\boldsymbol{\theta}|\mathbf{y}_{i,\,j_{\mathrm{tmp}}}) d\boldsymbol{\theta}} \tag{4.13}$$

denote the posterior density of $\boldsymbol{\theta}_i$ after j_{tmp} items. Then the Bayesian Credible Region (BCR) method would select the alternative hypothesis if $\int_{\Theta_m} \pi(\boldsymbol{\theta}|\mathbf{y}_{i,\,j_{\mathrm{tmp}}}) > 1 - \alpha$, the null hypothesis if $\int_{\Theta_n} \pi(\boldsymbol{\theta}|\mathbf{y}_{i,\,j_{\mathrm{tmp}}}) > 1 - \alpha$, and administer another item if neither of those conditions held. Because the ultimate decision relates to the posterior probability of mastery, one could think of the decision process supported by Berger (2012) depending on bounds of a Bayesian credible region.

4.2.3 Multidimensional Curtailed Procedures

As in the unidimensional case, multivariate versions of the SPRT (and GLR) assume potentially unlimited test lengths. One could also generalize curtailed methods to multidimensional classification algorithms. Because curtailed procedures only depend on the linked sequential decision procedure, these methods are easily generalized to multidimensional adaptive tests. Specifically, let $D_{i,\,j_{\text{tmp}}}$ be the temporary decision after $j_{\text{tmp}} < j_{\max}$, and assume that the test has not been stopped by j_{tmp} items. As in the unidimensional case, set $D_{j_{\text{tmp}}} = n$ if $C_{i,\,j_{\text{tmp}}} < (C_l + C_u)/2$, set $D_{j_{\text{tmp}}} = m$ if $C_{i,\,j_{\text{tmp}}} > (C_l + C_u)/2$, and pick two error rates, $0 \leq \epsilon_1 < .5$ and $0 \leq \epsilon_2 < .5$. Then the probability of being declared a non-mastery by maximum test length is

$$\Pr_{\tilde{\boldsymbol{\theta}}}(D_{i,\,j_{\max}} = n | C_{i,\,j_{\text{tmp}}}) = 1 - \Pr_{\tilde{\boldsymbol{\theta}}}(D_{i,\,j_{\max}} = m | C_{i,\,j_{\text{tmp}}}) \approx \Phi\left(\frac{C_0 - \mathbb{E}_{\tilde{\boldsymbol{\theta}}}(C_{i,\,j_{\max}} | C_{i,\,j_{\text{tmp}}})}{\sqrt{\text{Var}_{\tilde{\boldsymbol{\theta}}}(C_{i,\,j_{\max}} | C_{i,\,j_{\text{tmp}}})}}\right), \tag{4.14}$$

where

$$\mathbb{E}_{\tilde{\boldsymbol{\theta}}}(C_{i,\,j_{\max}} | C_{i,\,j_{\text{tmp}}}) = C_{i,\,j_{\text{tmp}}} + \sum_{j=j_{\text{tmp}}+1}^{j_{\max}} \mathbb{E}_{\tilde{\boldsymbol{\theta}}}\left(\log\left[\frac{L(\boldsymbol{\theta}_u | y_{ij})}{L(\boldsymbol{\theta}_l | y_{ij})}\right]\right), \tag{4.15}$$

$$\text{Var}_{\tilde{\boldsymbol{\theta}}}(C_{i,\,j_{\max}} | C_{i,\,j_{\text{tmp}}}) = \sum_{j=j_{\text{tmp}}+1}^{j_{\max}} \text{Var}_{\tilde{\boldsymbol{\theta}}}\left(\log\left[\frac{L(\boldsymbol{\theta}_u | y_{ij})}{L(\boldsymbol{\theta}_l | y_{ij})}\right]\right), \tag{4.16}$$

$\tilde{\boldsymbol{\theta}}$ is the assumed ability under which the expectation/variance are evaluated, and $\Phi(\cdot)$ is the CDF of a standard normal distribution. Although the expectation and variance are with respect to a vector, the probability, and thus, the likelihood ratio, are scalar functions. In essence, the multivariate SPRT with Stochastic Curtailment (M-SCSPRT) results in a unidimensional SCSPRT where the direction of projection is normal to the nearest point on the classification bound function (however defined).

Given the simple generalization of SCSPRT to multiple dimensions, the multivariate SPRT with Predictive Power (M-PPSPRT) can then be determined using

$$\Pr_{\boldsymbol{\Theta}}(D_{i,\,j_{\max}} = n | C_{i,\,j_{\text{tmp}}}) = \int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}|\mathbf{y}_{i,\,j_{\text{tmp}}}) \Pr_{\boldsymbol{\theta}}(D_{i,\,j_{\max}} = n | C_{i,\,j_{\text{tmp}}}) d\boldsymbol{\theta}, \tag{4.17}$$

where $\pi(\boldsymbol{\theta}|\mathbf{y}_{i,\,j_{\text{tmp}}})$ is the posterior distribution of $\boldsymbol{\theta}$ given response pattern $\mathbf{y}_{i,\,j_{\text{tmp}}}$ and prior distribution $\pi(\boldsymbol{\theta})$, as defined in Equation (4.13). The M-PPSPRT takes every possible $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, projects each point onto the nearest part of the classification bound function, uses the vector normal to the tangent plane for that part of the classification bound function to calculate the SCSPRT, and weights each SCSPRT by the corresponding $\boldsymbol{\theta}$'s posterior density. In both the multidimensional SCSPRT (M-SCSPRT) and multidimensional PPSPRT (M-PPSPRT), practitioners would compare the probability of non-mastery by the end of the test to $1 - \epsilon_1$ if $D_{j_{\text{tmp}}} = n$ or to ϵ_2 if $D_{j_{\text{tmp}}} = m$.

In the current section, I have proposed several novel stopping rules for multidimensional mastery testing, including the: (1) Constrained SPRT (C-SPRT), (2) Projected SPRT (P-SPRT), (3) Multidimensional GLR (M-GLR), (4) Weighted Likelihood Ratio (W-GLR), (5) Bayesian Credible Region (BCR), (6) Multidimensional SCSPRT (M-SCSPRT), and (7) Multidimensional PPSPRT (M-PPSPRT). In the next chapter, I propose a study to assess several of these stopping rules (in terms of classification accuracy and average test length) when implementing multidimensional mastery tests. However, each of these stopping rules requires a method of selecting future items items. As in the unidimensional case, the method of selecting items is integral to the performance of M-SCSPRT and M-PPSPRT algorithms. Therefore, in the final section of this chapter, I describe algorithms appropriate for choosing items for multidimensional classification tests.

4.3 Multidimensional Item Selection Algorithms

Multidimensional mastery testing algorithms also require methods of selecting items. Most of the unidimensional item selection algorithms, including those based on Fisher information and Kullback-Leibler divergence, have been generalized to multidimensional adaptive tests. The current section details common item selection algorithms in multidimensional adaptive tests and proposes novel item selection algorithms appropriate for mastery problems.

4.3.1 Fisher Information Methods

Many of the common Fisher information-based algorithms using in multidimensional adaptive tests were summarized by Frey and Seitz (2009). In general, Fisher information is defined as the negative expected second derivative of the log-likelihood with respect to $\boldsymbol{\theta}$. With respect to the 3PL C-MIRT model, Fisher information for item j can be written as a function of true $\boldsymbol{\theta}$ (e.g., Wang & Chang, 2011),

$$\mathcal{I}_j(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 \log[L(\boldsymbol{\theta}|\mathbf{y})]}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right] = \frac{[1-p_j(\boldsymbol{\theta})][p_j(\boldsymbol{\theta})-c_j]^2}{p_j(\boldsymbol{\theta})[1-c_j]^2}\mathbf{a}\mathbf{a}^T, \tag{4.18}$$

where $p_j(\boldsymbol{\theta})$ is defined in Equation (4.1). As in the unidimensional 3PL, Fisher information relates to the asymptotic variance of $\hat{\boldsymbol{\theta}}$ for a given $\boldsymbol{\theta}$. Unfortunately, Equation (4.18) is a matrix, so that choosing an item to minimize the asymptotic variance of $\hat{\theta}_1$ does not necessarily contribute to minimizing the asymptotic variance of any other $\hat{\theta}_k$ (where $k > 1$). Several methods have been proposed to aggregate information across all of the dimensions to choose succeeding items. One commonly used method (called the D-Method) chooses items that minimize the volume of the confidence ellipsoid for $\boldsymbol{\theta}$ (e.g., Segall, 1996). Assume that $\boldsymbol{\theta}$ is normally distributed with prior variance $\boldsymbol{\Sigma}$. Then the volume of a confidence ellipsoid for $\boldsymbol{\theta}$ after item j_{tmp} is proportional to

$$\text{GVar}(\boldsymbol{\theta}) = \left| \sum_{j=1}^{j_{\text{tmp}}-1} \mathcal{I}_j(\boldsymbol{\theta}) + \mathcal{I}_{j_{\text{tmp}}}(\boldsymbol{\theta}) + \boldsymbol{\Sigma}^{-1} \right|^{-1}. \tag{4.19}$$

If the first $j_{\text{tmp}} - 1$ items have already been administered, then practitioners can choose the item j_{tmp} that minimizes Equation (4.19) at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_i$.

Another commonly method of aggregating information of the Fisher information matrix (called the T-Method) chooses the next item by minimizing the average asymptotic variance of $\hat{\boldsymbol{\theta}}_i$ across all dimensions. Because the total variance is simply the sum of asymptotic variances, practitioners can choose item j_{tmp} by minimizing

$$\text{TVar}(\hat{\boldsymbol{\theta}}_i) = \text{tr}\left[\left(\sum_{j=1}^{j_{\text{tmp}}-1} \mathcal{I}_j(\hat{\boldsymbol{\theta}}_i) + \mathcal{I}_{j_{\text{tmp}}}(\hat{\boldsymbol{\theta}}_i) + \boldsymbol{\Sigma}^{-1} \right)^{-1} \right], \tag{4.20}$$

where the trace function, $\text{tr}(\cdot)$, adds together all of the elements on the diagonal of the matrix inside.

van der Linden (1999) suggested selecting items to minimize the variance in a particular direction (called the L-Method). Following van der Linden (1999), assume that a practitioner wants to summarize ability in multiple directions by using the composite score $\boldsymbol{\lambda}^T\boldsymbol{\theta}$, where $\boldsymbol{\lambda}$ is a vector of positive numbers that sum to 1. Then the variance of $\boldsymbol{\lambda}^T\hat{\boldsymbol{\theta}}_i$ after item j_{tmp} can be written as

$$\text{Var}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\theta}}_i) = \boldsymbol{\lambda}^T\text{Var}(\hat{\boldsymbol{\theta}}_i)\boldsymbol{\lambda} = \boldsymbol{\lambda}^T\left[\sum_{j=1}^{j_{\text{tmp}}-1} \mathcal{I}_j(\hat{\boldsymbol{\theta}}_i) + \mathcal{I}_{j_{\text{tmp}}}(\hat{\boldsymbol{\theta}}_i) + \boldsymbol{\Sigma}^{-1} \right]^{-1}\boldsymbol{\lambda}. \tag{4.21}$$

All of the proposed Fisher information algorithms can be varied by taking weighted averages of these functions across a well-defined region to account for variability in the maximum likelihood estimate. Moreover, all of the unidimensional criticisms of maximizing Fisher information at $\hat{\theta}_i$ also apply to minimizing one of the asymptotic

variances defined by Equations (4.19)–(4.21). In fact, Reckase (2009) explained that "having both a correct and incorrect response [is not] sufficient to guarantee that the location of the maximum likelihood point [for a multivariate ability vector] would have finite values for all coordinates" (p. 142). Unlike the unidimensional case, obtaining a finite maximum likelihood estimate for $\boldsymbol{\theta}_i$ using the C-MIRT model requires correct and incorrect responses on a variety of items that measure different dimensions of the underlying space. Additionally, selecting items by minimizing Equations (4.19)–(4.21) at $\hat{\boldsymbol{\theta}}_i$ is still not optimal in determining whether $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}_m$. Therefore, I will shortly propose modified item selection algorithms appropriate for multidimensional mastery testing. First, I explain alternative multidimensional item selection algorithms, based on Kullback Leibler divergence, that directly account for the uncertainty in $\hat{\boldsymbol{\theta}}_i$ early in an adaptive test.

4.3.2 Kullback-Leibler Methods

Kullback-Leibler divergence methods are generally recommended to account for the uncertainty in $\hat{\boldsymbol{\theta}}_i$ early in an adaptive test. Veldkamp and van der Linden (2002) generalized Chang and Ying's (1996) KL information index to multiple dimensions. As before, let the Kullback-Leibler (KL) divergence for the j^{th} item be defined as (see Section 2.3.2 for corresponding motivation)

$$\text{KL}_j(\boldsymbol{\theta}_i||\boldsymbol{\theta}) = p_j(\boldsymbol{\theta}_i) \log \left[\frac{p_j(\boldsymbol{\theta}_i)}{p_j(\boldsymbol{\theta})} \right] + [1 - p_j(\boldsymbol{\theta}_i)] \log \left[\frac{1 - p_j(\boldsymbol{\theta}_i)}{1 - p_j(\boldsymbol{\theta})} \right], \tag{4.22}$$

where $p_j(\boldsymbol{\theta}_i)$ is determined by the C-MIRT model as defined in Equation (4.1). Then the multidimensional KL divergence index can be defined as

$$\text{KL}_j(\hat{\boldsymbol{\theta}}_i|w_{ij}) = \int_{\boldsymbol{\Theta}_D} w_{ij} \text{KL}_j(\hat{\boldsymbol{\theta}}_i||\boldsymbol{\theta}) \mu(d\boldsymbol{\theta}), \tag{4.23}$$

where w_{ij} is some weight function, usually the prior distribution of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$, or the posterior distribution of $\boldsymbol{\theta}_i$ after $j-1$ items, $\pi(\boldsymbol{\theta}|\mathbf{y}_{i,\,j-1})$. In Equation (4.23), one must specify a (generally symmetric) domain of integration, which is represented by $\boldsymbol{\Theta}_D$. Wang and Chang (2011) proposed that $\boldsymbol{\Theta}_D$ in two-dimensions be square, rectangular, circular, or elliptical with $\hat{\boldsymbol{\theta}}_i$ as the center of the domain. In three (or more dimensions), $\boldsymbol{\Theta}_D$ would usually be a (hyper) cube, rectangular prism, sphere, or ellipsoid, although one could justify using additional geometric shapes given a particular problem. Oddly, Wang and Chang (2011) found that KL divergence indices resulted in a larger Euclidean distance between $\hat{\boldsymbol{\theta}}_i$ and $\boldsymbol{\theta}_i$ than D-Method Fisher information and multidimensional mutual information indices, which is inconsistent with comparable unidimensional algorithms. They claimed that "in the multidimensional case, items with larger [KL divergence] do not necessarily provide higher power for discriminating θ_1 from $\hat{\theta}_1$" (p. 379), and suggested that "perhaps, this phenomenon can be further boiled down to the compensatory nature of the [C-MIRT] model" (p. 380). As will be shown forthwith, the Fisher and KL divergence indices can be applied to mastery problems with a well-specified domain of integration. In fact, the corresponding domain of integration for multidimensional mastery problems might alleviate inefficiencies in the multidimensional KL divergence algorithm.

4.3.3 Mastery Testing Methods

Based on Chapter 3, selecting items by maximizing information at the classification bound (in unidimensional CMT) typically results in shorter average tests than selecting items by maximizing information at proximate ability estimates. One would suspect that conclusions drawn from Chapter 3 should carry over into multidimensional mastery testing algorithms. Therefore, selecting items at the classification bound function

separating mastery from non-mastery should result in shorter average tests than selecting items at $\hat{\boldsymbol{\theta}}_i$. Unfortunately, any attempt to generalize cut-point based algorithms to multidimensional mastery testing leads to a familiar problem. How does one choose an item based on the cut-point separating categories when that cut-point is an uncountably infinite set? Not surprisingly, in parallel with the solutions described for generalizations of the SPRT, one finds several, reasonable solutions.

Let $\hat{\boldsymbol{\theta}}_0$ be the optimal point on the classification bound function as defined by either P-SPRT or C-SPRT. Then a simple cut-point based item selection algorithm would select items that minimize Equation (4.19) or Equation (4.20) at $\hat{\boldsymbol{\theta}}_0$. However, practitioners generally desire items that separate masters from non-masters (along the line perpendicular to the classification bound function), and Equations (4.19) and (4.20) consider all directions as equally important. A more sophisticated algorithm would find the line normal to the tangent plane defined by $\hat{\boldsymbol{\theta}}_0$, $\boldsymbol{\theta}_\delta = \frac{\nabla g(\hat{\boldsymbol{\theta}}_0)}{||\nabla g(\hat{\boldsymbol{\theta}}_0)||}$, and then select items that minimize Equation (4.21) with $\boldsymbol{\lambda} = \boldsymbol{\theta}_\delta$.

Unlike the unidimensional case, choosing the optimal cut-point for multidimensional mastery testing requires a (fallible) estimate of $\hat{\boldsymbol{\theta}}_0$. One could better account for uncertainty in $\hat{\boldsymbol{\theta}}_0$ by maximizing information (or minimizing the asymptotic variance) across a region. With respect to multidimensional Fisher information, one could take Equation (4.21) with $\boldsymbol{\lambda} = \boldsymbol{\theta}_\delta$ and average across the surface defined by $g(\boldsymbol{\theta}) = 0$. In other words, the proposed algorithm would be based on integrating $f_j(\boldsymbol{\theta}|w_{ij}) = \frac{w_{ij}}{||\nabla g(\boldsymbol{\theta})||^2} \nabla g(\boldsymbol{\theta})^T \text{Var}(\boldsymbol{\theta}) \nabla g(\boldsymbol{\theta})$ along the surface defined by $g(\boldsymbol{\theta}) = 0$. For simplicity, assume that $K \in \{2, 3\}$ and a parameterization of the surface can either be written as $\mathbf{r}(\theta_1) = [\theta_1,\, f_s(\theta_1)]$ or $\mathbf{r}(\boldsymbol{\theta}) = [\theta_1,\, \theta_2,\, f_s(\theta_1,\, \theta_2)]$. Then Fisher information averaged across the classification bound function/surface (S-FI) would be defined as

$$S\mathcal{I}_j(\boldsymbol{\theta}|w_{ij}) = \int_{\Theta_1} f(\mathbf{r}(\theta_1)|w_{ij}) \left|\left| \frac{\mathbf{r}(\theta_1)}{d\theta_1} \right|\right| d\theta_1 \tag{4.24}$$

if $K = 2$ or

$$S\mathcal{I}_j(\boldsymbol{\theta}) = \int_{\boldsymbol{\Theta}_D} f(\mathbf{r}(\boldsymbol{\theta})|w_{ij})) \left|\left| \frac{\mathbf{r}(\boldsymbol{\theta})}{d\theta_1} \times \frac{\mathbf{r}(\boldsymbol{\theta})}{d\theta_2} \right|\right| d\theta_1 d\theta_2 \tag{4.25}$$

if $K = 3$, where $\times$ stands for the cross-product operator, $\boldsymbol{\Theta}_D$ denotes the area of integration in $\Theta_1 \times \Theta_2$, and w_{ij} is either the prior distribution of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$, or the posterior distribution of $\boldsymbol{\theta}_i$ after $j-1$ items, $\pi(\boldsymbol{\theta}|\mathbf{y}_{i,\,j-1})$. Note that the above parameterization cannot be used for non-compensatory classification bound functions. Instead, parameterize $\mathbf{r}$ in terms of an auxiliary variable, t, where $\mathbf{r}(t)$ is on the line $[\theta_{01} - t,\, \theta_{02}]$ when $t < 0$ and the line $[\theta_{01},\, \theta_{02} + t]$ when $t > 0$. Due to the lack of a derivative for $t = 0$, $T = (-\infty, 0) \times (0, \infty)$. A similar $\mathbf{r}(t_1,\, t_2)$ parameterization should be specified for a three-dimensional, non-compensatory, classification bound function.

All of the heretofore mentioned multidimensional mastery testing item selection algorithms are based on maximizing Fisher information (or a function of Fisher information) at (or along) the classification bound function. Not surprisingly, the methods just described also apply to the other item selection algorithms with minimal alterations. To construct a KL divergence based item selection algorithm for multidimensional mastery testing, let $\boldsymbol{\theta}_u = \boldsymbol{\theta} + \delta \nabla g(\boldsymbol{\theta})$ and $\boldsymbol{\theta}_l = \boldsymbol{\theta} - \delta \nabla g(\boldsymbol{\theta})$ be vectors of points normal to the classification bound function at some $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. Then one could define $f_j(\mathbf{r}(\boldsymbol{\theta})|w_{ij}) = w_{ij}\mathrm{KL}_j(\boldsymbol{\theta}_u||\boldsymbol{\theta}_l)$ and evaluate $f(\mathbf{r}(\boldsymbol{\theta})|w_{ij})$ at $\hat{\boldsymbol{\theta}}_0$ (similar to L-FI at the classification bound function). Alternatively, one could integrate $f_j(\mathbf{r}(\boldsymbol{\theta})|w_{ij})$ between endpoints equidistant from $\hat{\boldsymbol{\theta}}_0$.

As in the unidimensional case, the proposed KL divergence index for multidimensional mastery testing assumes that every examinee is in the mastery region. One could also generalize the weighted log-odds ratio (LO; Lin & Spray, 2000) or mutual information (MI; Weissman, 2007) item selection algorithms to multidimensional classification

problems. To generalize the weighted log-odds ratio, let

$$\mathrm{LO}_j(\boldsymbol{\theta}_u||\boldsymbol{\theta}_l) = \sum_y \mathbb{E}\log\left(\left[\frac{p_j(\boldsymbol{\theta}_u)}{p_j(\boldsymbol{\theta}_l)}\right]^Y \div \left[\frac{1-p_j(\boldsymbol{\theta}_u)}{1-p_j(\theta_l)}\right]^{1-Y}\right), \tag{4.26}$$

$$= \mathbb{E}(Y=1)\log\left[\frac{p_j(\boldsymbol{\theta}_u)}{p_j(\boldsymbol{\theta}_l)}\right] - [1-\mathbb{E}(Y=1)]\log\left[\frac{1-p_j(\boldsymbol{\theta}_u)}{1-p_j(\boldsymbol{\theta}_l)}\right] \tag{4.27}$$

where $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_l$ are defined above and $\mathbb{E}(Y=1)$ is the classical difficulty of an item as explained in Chapter 2. Yet Chapter 3 indicated that the optimal item for an SPRT algorithm depends on the location of true ability relative to the classification bound, and Equation (4.26) only considers the location of the average examinee. A better algorithm would take the expected log-likelihood (as in the KL divergence index) conditional on proximate ability estimates. This expected log-likelihood ratio function for the j^{th} item would then be

$$\mathrm{ELR}_j(\hat{\boldsymbol{\theta}}_i) = \mathbb{E}_{\hat{\boldsymbol{\theta}}_i}\left[\log\left[\mathrm{LR}(\boldsymbol{\theta}_u, \boldsymbol{\theta}_l|Y_{ij})\right]\right] \tag{4.28}$$

$$= p_j(\hat{\boldsymbol{\theta}}_i)\log\left[\frac{p_j(\boldsymbol{\theta}_u)}{p_j(\boldsymbol{\theta}_l}\right] + [1-p_j(\hat{\boldsymbol{\theta}}_i)]\log\left[\frac{1-p_j(\boldsymbol{\theta}_u)}{1-p_j(\boldsymbol{\theta}_l)}\right], \tag{4.29}$$

where $\mathrm{ELR}_j(\hat{\boldsymbol{\theta}}_i)$ stands for "the expected log-likelihood ratio for prospective item j given $\hat{\boldsymbol{\theta}}_i$", and $p_j(\hat{\boldsymbol{\theta}}_i)$ is calculated using Equation (4.1) with $\hat{\boldsymbol{\theta}}_i$ inserted in place of $\boldsymbol{\theta}_i$. Finally, if $\hat{\boldsymbol{\theta}}_i \in \boldsymbol{\Theta}_m$, then item j should be chosen to maximize Equation (4.29), whereas if $\hat{\boldsymbol{\theta}}_i \in \boldsymbol{\Theta}_n$, then item j should be chosen to minimize Equation (4.29).

Mulder and van der Linden (2010) and Wang and Chang (2011) also generalized a mutual information item selection rule to multiple dimensions. In general, mutual information, $\mathrm{MI}(x;\, y) = \sum_x \sum_y f(x,\, y)\log\left[\frac{f(x,y)}{f(x)f(y)}\right]$, is the KL divergence between the joint distribution of $(x,\, y)$ and the product of the marginal distributions. Mulder and

van der Linden (2010) took $f(y)$ to be $\pi(\boldsymbol{\theta}|\mathbf{y}_{j-1})$, the posterior distribution of $\boldsymbol{\theta}$ after $j-1$ items and took $f(x)$ to be $\Pr_j(Y = y|\mathbf{y}_{i,\,j-1})$, the posterior predictive distribution given the previous responses. Then multidimensional mutual information simplifies to

$$\mathrm{MI}_j(\boldsymbol{\Theta}_D) = \sum_y \int_{\boldsymbol{\Theta}_D} \Pr_j(Y = y|\mathbf{y}_{i,\,j-1})\pi(\boldsymbol{\theta}|\mathbf{y}_{j-1}) \log \left[\frac{\Pr_j(Y = y|\boldsymbol{\theta})}{\Pr_j(Y = y|\mathbf{y}_{i,\,j-1})}\right] d\boldsymbol{\theta}. \quad (4.30)$$

where $\boldsymbol{\Theta}_D$ is the domain of integration. Because Mulder and van der Linden (2010) and Wang and Chang (2011) were assessing multidimensional item selection rules in precision-based CAT, $\boldsymbol{\Theta}_D$ was taken to be the entire space. However, a more appropriate item selection rule in MCMT would take the surface integral of $f(\mathbf{r}(\boldsymbol{\theta})|\mathbf{y}_{i,\,j-1}) = \sum_y \Pr_j(y = y|\mathbf{y}_{i,\,j-1})\pi(\boldsymbol{\theta}|\mathbf{y}_{j-1}) \log \left[\frac{\Pr_j(Y=y|\boldsymbol{\theta})}{\Pr_j(Y=y|\mathbf{y}_{i,\,j-1})}\right]$ across $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ such that $g(\boldsymbol{\theta}) = 0$.

I therefore have described several item selection algorithms appropriate for multidimensional mastery testing, including: (1) D-FI at $\hat{\boldsymbol{\theta}}_0$, (2) T-FI at $\hat{\boldsymbol{\theta}}_0$, (3) L-FI at $\hat{\boldsymbol{\theta}}_0$ with $\boldsymbol{\lambda} = \boldsymbol{\theta}_\delta$, (4) S-FI along $\boldsymbol{\Theta}_0$, (5) L-KL comparing $\boldsymbol{\theta}_u$ to $\boldsymbol{\theta}_l$, (6) S-KL along $\boldsymbol{\Theta}_0$, (7) L-LO comparing $\boldsymbol{\theta}_u$ to $\boldsymbol{\theta}_l$, (8) L-ELR comparing $\boldsymbol{\theta}_u$ to $\boldsymbol{\theta}_l$ conditional on $\hat{\boldsymbol{\theta}}_i$, and (9) M-MI. In the next chapter, I describe simulations intended to compare several of the proposed MCMT item selection algorithms and stopping rules given a variety of item bank and distributional configurations.

Chapter 5

Study Design and Procedures

In this chapter, I describe a simulation study that was designed to compare test length and classification accuracy for a variety of MCMT stopping rules and item selection algorithms. I first discuss properties of the latent trait distribution, IRT model, and testing process. I then outline the ability estimation methods, item selection algorithms, and stopping rules that were used in the simulations.

5.1 Assessment Properties

5.1.1 Item Bank and IRT Model

Two simulated item banks were employed, each consisting of $J = 900$ items on $K = 2$ dimensions with parameters calibrated according to the C-MIRT model, as defined in Equation (4.1). Following Wang and Chen (2004), one of the item banks was constructed to have between-item multidimensionality, and the other item bank was constructed to have within-item multidimensionality. Note that Reckase (2009) acknowledged that little work has adequately addressed necessary properties of multidimensional item banks. Therefore, the item banks were constructed to be similar in overall information under

the assumption that the relative pattern of results should apply to other pairs of similar item banks.

Both of the C-MIRT item parameter banks were constructed according to the following algorithm:

- MDISC-parameters were generated from a log-normal distribution with a log-mean of $\mu_{\log} = 0.50$ and a log-standard deviation $\sigma_{\log} = 0.10$ (corresponding to a mean of approximately $\mu = 1.657$ and a standard deviation of approximately $\sigma = 0.166$, and is similar to how the a-parameters were generated in Babcock & Weiss, 2009). If an item loaded on both dimensions, the square of the first a-parameter, $a_{j1_1}^2$, was generated uniformly between 0 and MDISC_j^2, and the square of the second a-parameter, a_{j2}^2, was set to $\text{MDISC}_j^2 - a_{jk_1}^2$. If an item only loaded on one dimension, the square of the a-parameter corresponding to that dimension was set to MDISC_j^2.

- b-parameters were generated from a uniform distribution between -3.5 and 3.5, a distribution slightly wider than one adopted in Wang and Chen (2004). After generating b_j for an item, the corresponding threshold parameter was set to $d_j = -b_j \mathbf{a}_j^T \mathbf{1}$. Note that the C-MIRT model, as defined in Equation (4.1), is specified in terms of item threshold parameters rather than item difficulties.

- c-parameters were set to .2 to mimic a multiple choice test with five response possibilities per item.

As described by Wang and Chen (2004), item banks with within-item multidimensionality have individual items loading on more than one dimension. For this study, the within-item multidimensional bank had all $J = 900$ items loading on all $K = 2$ dimensions. Although this structure never could be recovered by a factor rotation matrix,

many of the items had near zero loadings on one of the dimensions, and psychometricians could conceivably set that loading to zero. In contrast to within-item multidimensional banks, the between-item multidimensional bank had individual items loading on only one dimension but the pooled items loading on multiple dimensions. One could think of between-item multidimensionality as an extreme form of simple structure where individual items tap only one part of a test. For this study, the between-item multidimensional bank had $J_k = 900/2 = 450$ items loading on each of the $K = 2$ dimensions.

5.1.2 Latent Trait Distribution

Two simulation studies were performed. First, the overall accuracy and average test length was estimated by simulating $N = 5{,}000$ $\boldsymbol{\theta}$s from a multivariate normal distribution with $\mu = \mathbf{0}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. Given the item banks described in the previous sub-section, three correlations were assumed between the traits: $\rho = 0$, $\rho = .33$, and $\rho = .67$. Second, conditional accuracy and test length were determined by simulating 1,000 MCMTs at each point along a 8×8 square where
$\theta_{ik} \in \{-.7, -.5, -.3, -.1, +.1, +.3, +.5, +.7\}$. The overall/aggregate test length and accuracy rates were determined from all combinations of conditions, but conditional performance (the second set of simulations) was only established for those conditions deemed efficient and highly accurate according to the distributional simulation. The exact conditions chosen for the conditional simulation as well as the reasons for choosing those conditions will be described in Chapter 6.

5.1.3 Classification Bound Functions

Two classification bound functions were used in the simulation: a linear, compensatory function where $\theta_2 = -\theta_1$ and a non-compensatory function with mastery defined as the first quadrant in Cartesian space. These classification bound functions are located

near the region of highest simulee density. Therefore, the classification tasks proposed in this study are difficult relative to tasks using other classification bound functions of a similar shape.

5.1.4 Overall CAT Algorithm

I imposed several restrictions on how items were administered to simulees across all conditions. Simulees were allowed to take between $J = 10$ and $J = 100$ items. The first four items were always randomly selected. Thereafter, the CAT algorithm bounced back and forth between estimating examinee $\boldsymbol{\theta}$, checking the stopping rule, and administering items until either the chosen stopping rule criterion had been satisfied or the maximum test length had been reached. These procedures are similar to those used in Nydick (2012). All simulations were performed in `R` (R Core Team, 2013) using code modified from the `catIrt` package (Nydick, 2013).

5.2 Adaptive Testing Procedures

5.2.1 Ability Estimation Algorithms

Several conditions require estimates of the latent trait to either select items or make a classification decision. Regardless of condition, estimating the latent trait depended solely on the number of administered items. Before administering each of the first four items, $\boldsymbol{\theta}$ was estimated to be at a randomly chosen point on the $[-1, 1] \times [-1, 1]$ square. Thereafter, $\boldsymbol{\theta}$ was estimated using a modified Maximum Likelihood Estimation (MLE) algorithm. Reckase (2009) noticed that "the maximum likelihood estimation has the problem that finite estimates of coordinates may not exist when the number of items that has been administered is small" (p. 320). Therefore, if allowing unbounded MLE estimates, one would obtain the counterintuitive finding that simulees with true ability

far away from the classification bound function take longer to classify under some stopping rules than simulees with true ability closer to the classification bound. One could alleviate concern for unbounded maximum likelihood estimates using two strategies: (1) assuming bounds on the latent trait or (2) adopting an alternative estimation method.

A common alternative estimation method to MLE is Bayesian Modal Estimation (BME). Bayesian modal estimates maximize the posterior distribution. As long as the posterior is a proper distribution, the BME is then

$$\hat{\boldsymbol{\theta}}_{\text{BME},\,i} = \arg\max_{\boldsymbol{\theta}} \Big\{ \log\left[L(\boldsymbol{\theta}|\mathbf{y}_i)\right] + \log[\pi(\boldsymbol{\theta})] \Big\}, \tag{5.1}$$

where $L(\boldsymbol{\theta}|\mathbf{y}_i)$ is the likelihood function as defined by Equation (4.8), and $\pi(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$. Unfortunately, the computing time required to find the maximum of the posterior distribution was exorbitantly long, even when running conditions in parallel and writing much of the code in a C loop. Therefore, I decided to estimate ability using a modified MLE algorithm, such that

$$\hat{\boldsymbol{\theta}}_{\text{MLE},\,i} = \underset{\boldsymbol{\theta}\in[-4,4]\times[-4,4]}{\arg\max} \Big\{ \log\left[L(\boldsymbol{\theta}|\mathbf{y}_i)\right] \Big\}. \tag{5.2}$$

Because none of the stopping rules nor item selection algorithms consider the variance of $\hat{\boldsymbol{\theta}}$, this MLE formation should result in less conservative stopping rule decisions than an ability estimation procedure based on Bayesian methods.

5.2.2 Item Selection Algorithms

Five of the item selection algorithms were adopted in the simulation: D-FI, L-FI, L-KL, S-KL, and L-ELR. Due to the lessons from Chapter 3, all of the algorithms selected items based on the location of the classification bound function. Moreover, each of these algorithms was described in Chapter 4 and will not be explained. The item selection

algorithms were chosen for several reasons. First, I wanted to compare an algorithm that examines all directions (e.g., D-FI) to algorithms that consider only the direction of the classification bound function. Because D-FI notes only the location and not the functional form of the classification bound function, one would expect selecting items by maximizing D-FI to be much less efficient than the other item selection algorithms. Second, I wanted to test an algorithm based on surface information. In preliminary simulations, S-FI and S-KL resulted in similar measures of efficiency. However, S-FI took at least .15 of a second per item, whereas S-KL took less than .04 of a second per item. As in estimating ability using BME, selecting items based on S-FI would have resulted in an exorbitantly and impractically long running time given the number of items, simulees, and overall conditions.

5.2.3 Stopping Rules

The ultimate goal of the simulation was to compare the classification accuracy and average test length for several stopping rules under a variety of conditions. Five of the stopping rules were explored in the study: P-SPRT, C-SPRT, M-GLR, M-SCSPRT, and BCR. Because P-SPRT did not result in accurate adaptive tests (as will be explored in Chapter 6), M-SCSPRT was based on the C-SPRT formulation. Moreover, W-GLR and M-PPSPRT were not tested due to their complex integrals and the requisite long computing time.

The stopping rules adopted the following parameterizations. With respect to the log-likelihood ratio-based methods, $\alpha = \beta = .1$ and $\delta \in \{.15, .25\}$. If using M-SCSPRT, $\alpha = \beta = .1$, $\delta \in \{.15, .25\}$, and ϵ_1 and ϵ_2 were set to .05. The rationale for applying these values in simulation was discussed in Thompson (2010) (and adapted for Nydick, 2012). Namely, Thompson (2010) concluded that "nominal [percent classified correctly] had very little effect on observed [percent classified correctly]" (p. 9). In addition to

Thompson's conclusions, the values chosen were similar to those recommended for use in unidimensional mastery testing (e.g., Eggen, 2011; Finkelman 2008a; Lin, 2011; Wang & Huang, 2011). With respect to BCR, α was set to either .05 or .10. These values are similar to the specified nominal error rates for the log-likelihood ratio-based stopping rules.

5.2.4 Overall Conditions

As described in this chapter, I simulated a variety of multidimensional mastery tests using various item selection algorithms, stopping rules, and distributional properties. Ultimately, there were 3 (ability correlations) $\times$ 2 (classification bound functions) $\times$ 2 (item banks) $\times$ 5 (item selection algorithms) $\times$ 10 (stopping rules) = 600 overall conditions. The next chapter compares, in depth, each of those 600 conditions to determine the optimal multidimensional mastery testing design.

Chapter 6

Simulation Results

These results are summarized in two sections. I first present the overall test length, classification accuracy, and loss of using particular combinations of conditions when simulating from a distribution. The first set of simulations lead to specific, optimally performing conditions. I then describe the conditional classification, accuracy, and overall loss of using the optimally performing set of conditions for several true ability vectors. Both simulations address different mastery testing goals: (1) How well the item selection algorithms and stopping rules perform given a distribution of examinees; and (2) How well the item selection algorithms and stopping rules perform for specific simulees near the classification bound function.

6.1 Results 1: Aggregated across a Distribution

The first set of simulations examined the accuracy and test length of various stopping rules, item selection algorithms, and item banks across a distribution of simulees[1]. Many practitioners require classification algorithms to be efficient and accurate for all

[1] All of the tables and figures generated by using statistics aggregated across a distribution of simulees are presented in Appendices B and C, respectively

examinees regardless of the distance between a person's ability level and the closest point on the classification bound function. Therefore, algorithms that perform well for examinees near the classification bound function serve little use unless those algorithms also easily and accurately classify more distant examinees. To determine the overall classification accuracy and test length, 5,000 $\boldsymbol{\theta} \sim N(\mathbf{0}, \mathbf{I}_2)$ (where $\mathbf{I}_k$ is a k-dimensional identity matrix) were simulated, rotated so that θ_1 and θ_2 correlated a specific amount[2] (either .00, .33, or .67, as explained in Chapter 5), and then tested using each combination of conditions.

Figure 6.1 displays the average test length and classification accuracy aggregated over simulees within each of the item selection algorithm (top panels) and stopping rule (bottom panels) conditions. To construct Figure 6.1, results were averaged across the latent ability correlation and the item bank conditions. Note that the left side of Figure 6.1 presents results when using a compensatory classification bound function, whereas the right side of Figure 6.1 presents results when using a non-compensatory classification bound function.

Consider the top panels of Figure 6.1. As shown in the upper left panel, item selection does not result in appreciable differences in classification accuracy when using a compensatory classification bound function. Conversely, the points are a bit more scattered in the accuracy direction when using a non-compensatory classification bound function, as shown in the upper right panel of Figure 6.1. However, the pattern of accuracy results are similar for both the compensatory and non-compensatory classification bound functions: L-FI results in higher accuracy than L-ELR and L-KL (which lead to similar accuracy rates), and S-KL performs at least as well as L-FI in terms of accuracy. Only D-FI has a contrasting accuracy pattern relative to the other item

[2]Let $\boldsymbol{\Sigma}_k = \mathbf{V}_k \boldsymbol{\Lambda}_k \mathbf{V}_k^T$ be the eigendecomposition of desired population covariance matrix $\boldsymbol{\Sigma}_k$ with dimensionality k, and let $\boldsymbol{\theta} \sim N(\mathbf{0}_k, \mathbf{I}_k)$. Then $\tilde{\boldsymbol{\theta}} = \mathbf{V}_k \boldsymbol{\Lambda}_k^{1/2} \boldsymbol{\theta} \sim N(\mathbf{0}_k, \boldsymbol{\Sigma}_k)$.

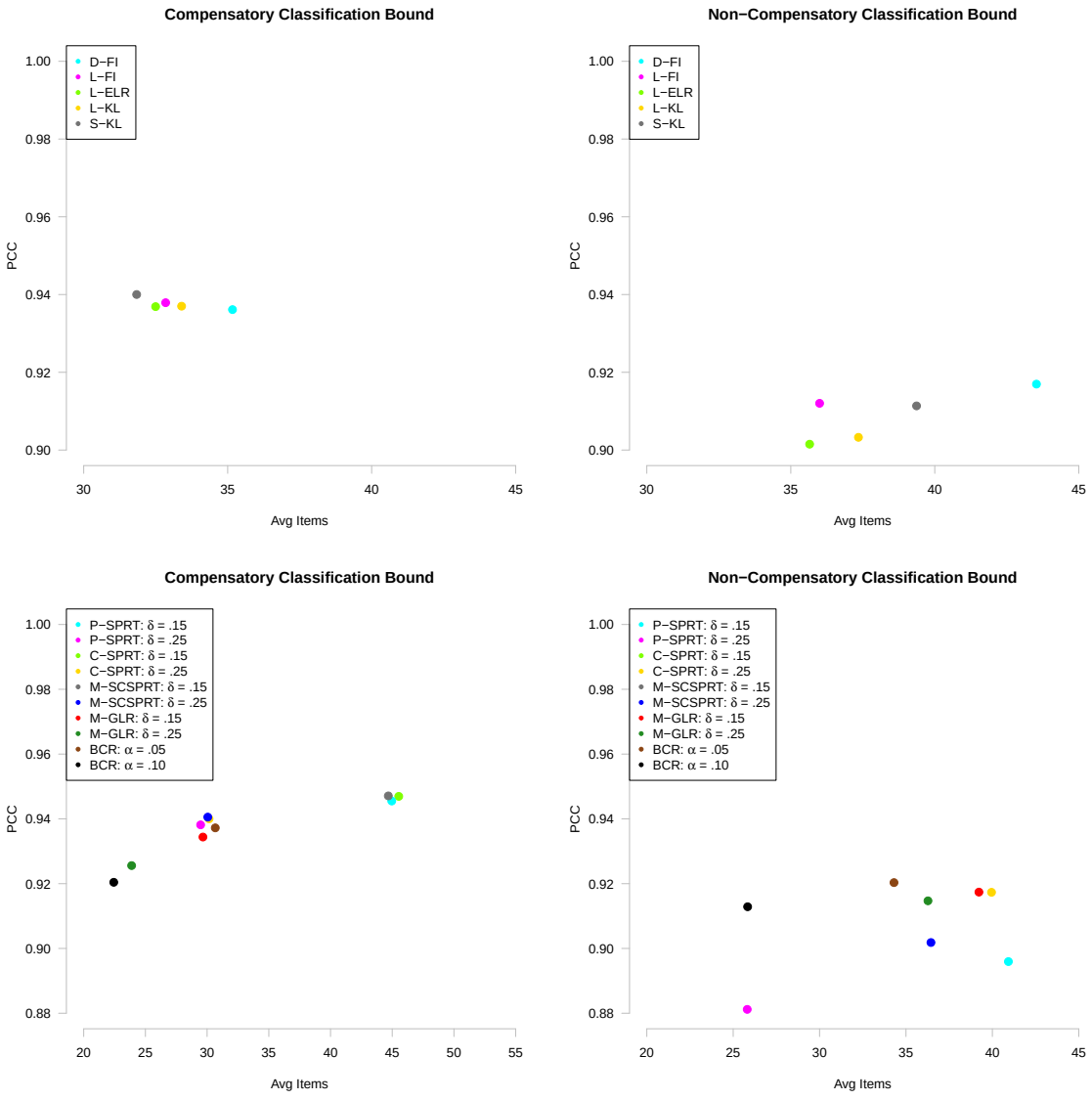

Figure 6.1: Scatterplots of the percent classified correctly (PCC) by average number of items administered for different item selection algorithms (top panel) and different stopping rules (bottom panels) using either a compensatory classification bound function (left panel) or a non-compensatory classification bound function (right panels).

selection algorithms when using a compensatory versus a non-compensatory classification bound function. When using a compensatory classification bound function, D-FI results in the lowest accuracy rate, but when using a non-compensatory classification bound function, D-FI results in the highest accuracy rate. This finding is probably due to the form of the classification bound function. D-FI chooses successive items based on maximizing the determinant of expected test information at the closest point on the classification bound function, whereas the other algorithms choose successive items by maximizing information in the direction normal to the classification bound function at that point. Because the non-compensatory classification bound function has two possible normal directions, D-FI might protect against items being chosen entirely along one of those directions. With respect to test length, S-KL leads to the shortest tests when using a compensatory classification bound but the second longest tests when using a non-compensatory classification bound. The other four item selection algorithms have a similar relative distance in average number of items regardless of classification bound function.

As described in an earlier section, several researchers (e.g., Finkelman, 2008b; 2010; Vos, 2000) have used a simple loss function to combine accuracy and test length into a single index. This specific loss function can be presented as

$$\text{Loss} = P \times I_W + J, \tag{6.1}$$

where I_W is an indicator function for incorrect classification, J is the number of items given to an examinee, and P is the penalty accrued for an incorrect decision. Because an average distributes over a linear function, one can write average loss as

$$\overline{\text{Loss}} = P \times (1 - \text{acc}) + \bar{J}, \tag{6.2}$$

where acc is the percent classified correctly and $\bar{J}$ is the average number of items across the appropriate distribution of simulees.

Figure 6.2 displays the average loss for those conditions represented in Figure 6.1 as the penalty accrued (P) goes from 0 – 3,000. So that relative loss can be compared as P increases, the calculated values of loss given a particular P were standardized across the relevant conditions at that P.

Consider the top panels of Figure 6.2 in parallel with the top panels of Figure 6.1. When using a compensatory classification bound function, the loss lines are nearly parallel. Surface KL divergence (S-KL) has the best loss, determinant Fisher information (D-FI) has the worst loss, and the remaining conditions are similar to each other irrespective of P. Due to the standardization, the loss plot magnifies very small differences in accuracy and test length across item selection algorithms when using a compensatory classification bound function. In contrast to the compensatory classification bound function, item selection algorithms have a different relationship between accuracy and test length when using a non-compensatory classification bound function. Linear Fisher information (L-FI) has the best loss until $P \approx 1{,}500$. Conversely, linear expected likelihood ratio (L-ELR) and L-KL result in the worst loss if weighting incorrect classifications as $P \geq 500$ but perform much better if $P \leq 300$. S-KL appears to protect against uninformative items better than D-FI and against poor classification decisions better than L-ELR and L-KL. Yet L-FI outperforms S-KL for all values of P. Therefore, the protection gained by using surface information does not appear to exceed that from items chosen in standard, well-selected directions.

Unlike item selection algorithms, varying the stopping rules appears to have a noticeable effect on classification accuracy, as shown on the bottom of Figure 6.1. Consider the bottom left panel of Figure 6.1. When using a compensatory classification bound

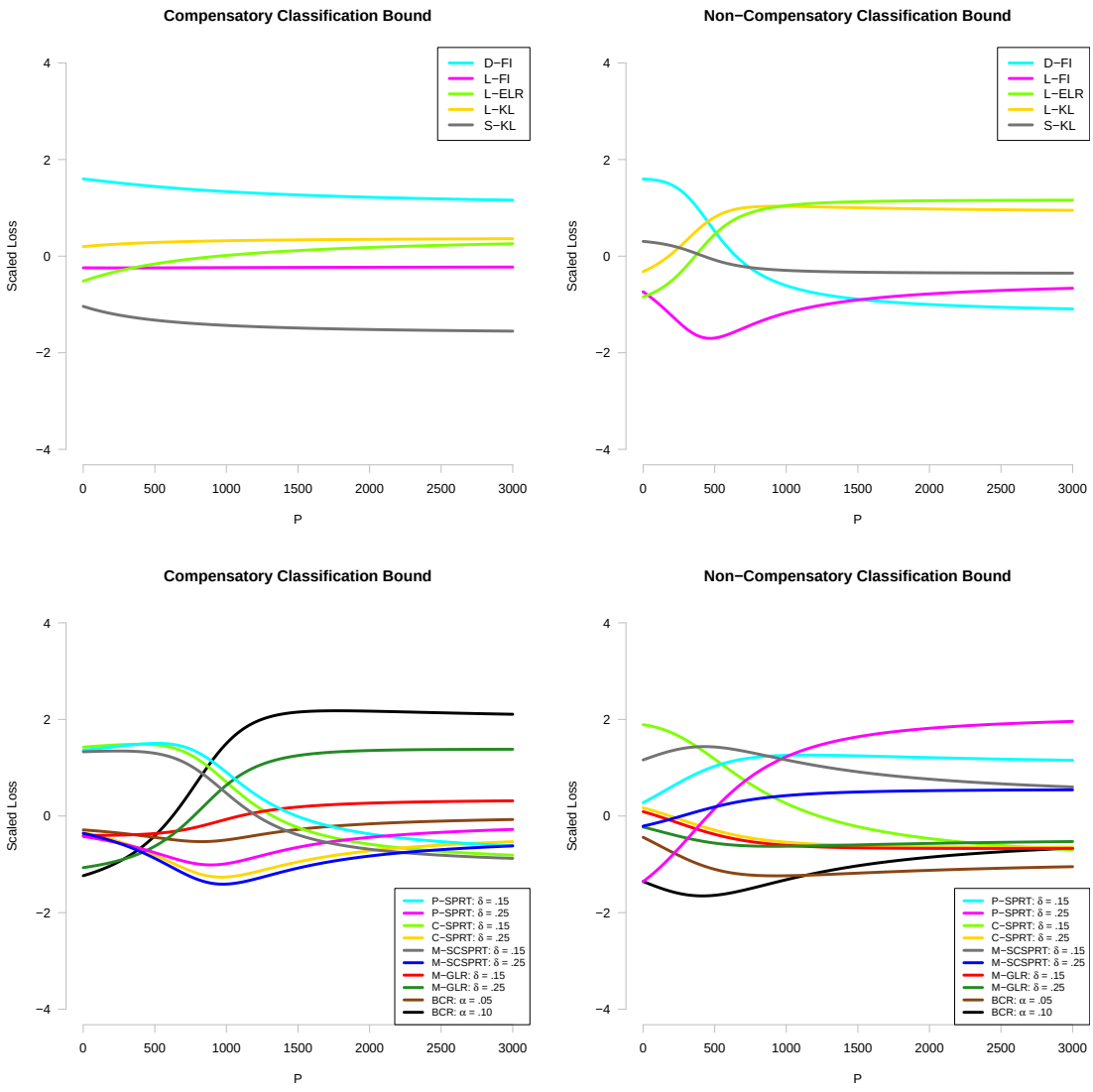

Figure 6.2: Average loss within each item selection algorithm or stopping rule for various values of P, where Loss $= P \times I_W + J$ (see Appendix B). The upper panels indicate the average loss for each of the item selection algorithms, whereas the lower panels indicate the average loss for each of the stopping rules. The left panels represent a compensatory classification bound function, whereas the right panels represent a non-compensatory classification bound function.

function, one finds a nearly linear relationship between number of items and classification accuracy. In fact, the OLS slope predicting accuracy from test length when using a compensatory classification bound function ($\hat{\beta}_1 = .0009$) is similar to the slope that I had earlier found in the unidimensional case with a test length of at least 17 items given comparable stopping rules ($\hat{\beta}_1 = .00085$; Nydick, 2012, p. 41). However, the OLS model yields a multiple R^2 of .774, indicating that average test length accounts for much of the variability in classification accuracy. Therefore, most of the stopping rules perform similarly when using a compensatory classification bound function, so that applying a slightly more conservative rule results in a slightly better classification accuracy rate. In contrast, applying various stopping rules to a classification test with a non-compensatory classification bound function leads to one of two consequences, as shown on the bottom right of Figure 6.1. First, a stopping rule could result in much worse classification accuracy compared to other stopping rules that lead to the same average number of administered items. For instance, P-SPRT with $\delta = .25$ leads to a similar average number of items as BCR with $\alpha = .05$ but results in a nearly .04 worse classification accuracy. Second, the stopping rule could result in a differing number of administered items without much changing the classification accuracy. For instance, after removing the four poorly performing conditions (i.e., the P-SPRT and M-SCSPRT conditions), the OLS slope predicting accuracy from test length when using a non-compensatory classification bound function is $\hat{\beta}_1 = .0003$ (with a slightly smaller R^2 of .68).

One could also examine various values of loss to better assess the overall performance of each stopping rule. These loss values are plotted on the bottom panels of Figure 6.2. Notice that when using the compensatory classification bound, scaled loss leads to 2-3 clumps of conditions as P increases. If $P \lessapprox 500$, then those conditions resulting in the shortest tests, such as BCR with $\alpha = .10$ and M-GLR with $\delta = .25$, also yield the

best loss. In fact, if $P \lessapprox 500$, then conditions cluster according to test length, and if $P \gtrapprox 1{,}000$, then conditions cluster according to accuracy. Note that C-SPRT and M-SCSPRT, both with $\delta = .25$, result in the least loss for the largest stretch of P and, thus, potentially optimally balance accuracy and test length concerns. In contrast, when using a non-compensatory classification bound function, scaled loss tends to cluster into two groups if $P \gtrapprox 500$, as shown on the bottom right panel of Figure 6.2. The C-SPRT, BCR, and M-GLR conditions result in similar tradeoffs between test length and classification accuracy, whereas the P-SPRT and M-SCSPRT conditions typically result in the worst loss. A simple conclusion when examining the bottom panels of Figures 6.1 and 6.2 is that the projected SPRT, and stochastic SPRT methods do not generalize to complex classification bound functions. This finding is possibly due to the non-compensatory classification bound function not aligning with the contours of the likelihood function and the consequences thereof for several stopping rules.

Thus far, I have only considered the overall effects of different item selection algorithms and stopping rules on the accuracy and test length of MCMTs given certain classification bound functions. I also examined the correlation between latent ability dimensions as well as two different item banks. As shown in Appendices B and C, one finds that lower correlations result in worse performing algorithms (both in accuracy and test length) regardless of classification bound function and irrespective of whether one conditions on other variables. For instance, a nearly identical relationship holds between latent ability correlations, test length, and classification accuracy within each of the item selection algorithms or stopping rules regardless of whether using a compensatory or non-compensatory classification bound function: lower correlations result in longer and less accurate tests. A more interesting relationship exists, though, between item selection algorithms or stopping rules, the chosen item bank, and test quality measures, as shown in in Figures 6.3 – 6.5.

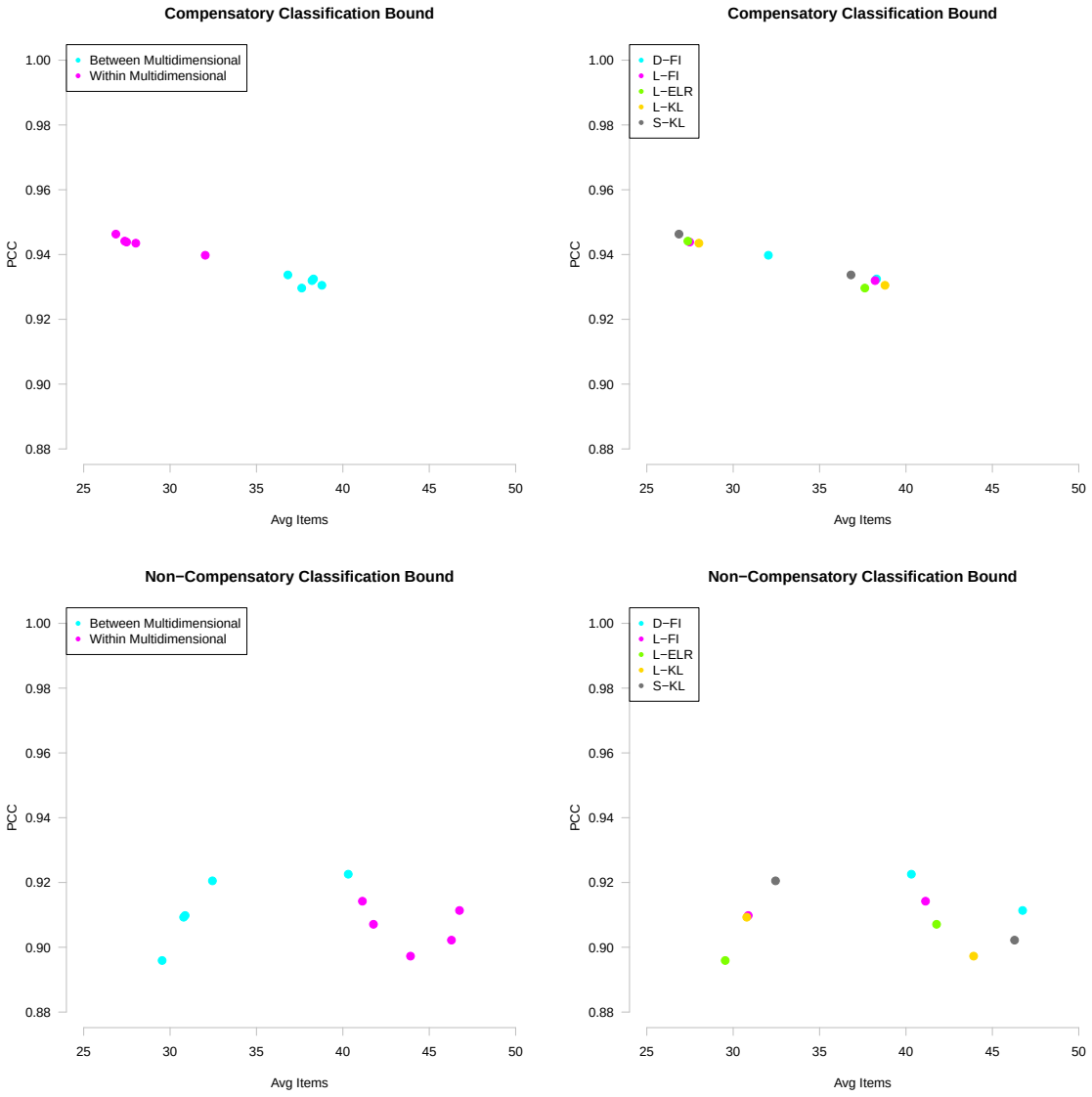

Figure 6.3: Scatterplots of the percent classified correctly (PCC) by average number of items administered based on the interaction between item bank and item selection algorithm using either a compensatory classification bound function (top panels) or a non-compensatory classification bound function (bottom panels). The left panels are color coded according to item bank, whereas the right panels are color coded according to item selection algorithm.

Figure 6.3 examines the relationship between item selection algorithm and average test length/classification accuracy within each of the item banks and conditioning on classification bound function. Consider the top panels of Figure 6.3, which examine this relationship given the compensatory classification bound function. Interestingly, one finds that when using a compensatory classification bound function, the within-item multidimensional bank results in the shortest and most accurate tests for all of the item selection algorithms. Every pink point is to the left of and above every blue point on the upper left quadrant of Figure 6.3. And within an item bank, one finds little difference between item selection algorithms in accuracy and test length but with a single exception. If using the compensatory classification bound function and the within-item multidimensional bank, D-FI results in much worse accuracy rates and much longer tests than the other item selection algorithms. With respect to the non-compensatory classification bound function, most of the item selection algorithms perform better when using a between-item multidimensional bank, as shown in the lower panels of Figure 6.3. In fact, all of the item selection algorithms result in shorter tests (and three of the five item selection algorithms result in more accurate tests) when using the between-item multidimensional bank than when using the within-item multidimensional bank. As when using the non-compensatory classification bound, only a single exception belies the general pattern of results: D-FI yields tests of similar length when adopting a between-item multidimensional bank to those conditions that use the within-item multidimensional bank. Therefore, D-FI performs similar to other item selection algorithms when using the worst item bank for a given classification problem but performs much worse than the other item selection algorithms when using the optimal item bank.

One also finds a clarifying representation of loss for various item selection algorithms when conditioning on different item banks, as shown in the upper panels of Figure 6.4.

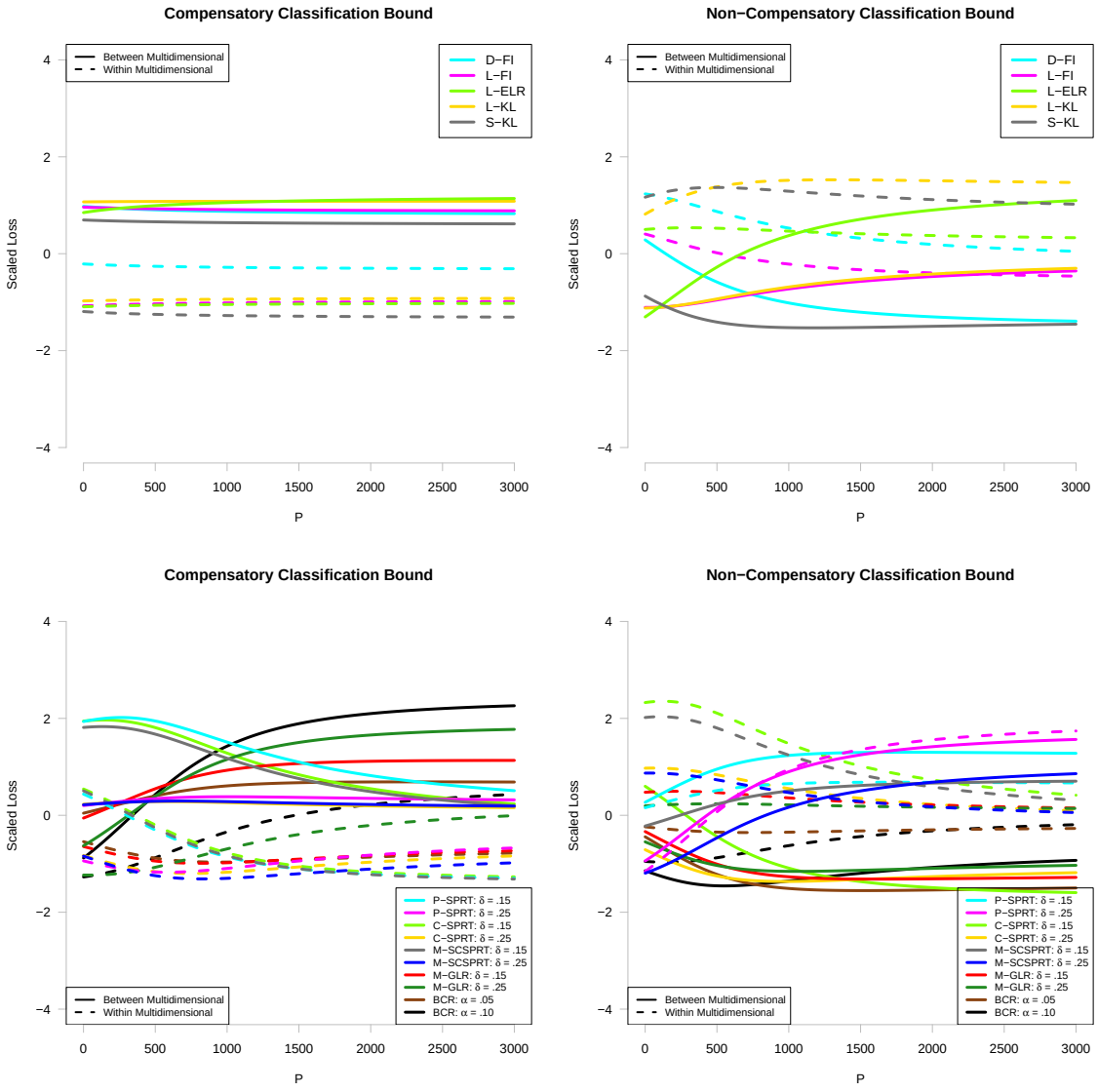

Figure 6.4: Average loss within each item selection algorithm by item bank or stopping rule by item bank for various values of P, where Loss $= P \times I_W + J$ (see Appendix B). The upper panels indicate the average loss for each of the item selection algorithms by item bank, whereas the lower panels indicate the average loss for each of the stopping rules by item bank. The left panels represent a compensatory classification bound function, whereas the right panels represent a non-compensatory classification bound function. Colors are coded according to item selection algorithm or stopping rule, whereas line type is determined by item bank.

Recall that if only examining item selection on loss given the non-compensatory classification bound function, L-FI results in the best balance between classification accuracy and test length for most values of loss, whereas S-KL results in worse loss than L-FI for all values of P. However, the earlier finding is entirely hidden by the poor performance of S-KL when using the inappropriate, within-item multidimensional bank. When using the appropriate item bank (within-item multidimensional for the compensatory classification bound and between-item multidimensional for the non-compensatory classification bound), S-KL yields the lowest loss for nearly all values of P, as evidenced by the dotted grey line on the upper left panel of Figure 6.4 and the solid grey line on the upper right panel of Figure 6.4 below all of the other lines on each respective panel. Moreover, with the sole exception of L-ELR, all of the non-compensatory classification bound item selection algorithms yield smaller relative loss until $P \approx 2{,}000$ when using the between-item multidimensional bank than when using the within-item multidimensional bank. As expected, one discovers few notable relationships between item selection and loss when using various item banks given the compensatory classification bound function. All of the lines on the upper right panel of Figure 6.4 are nearly parallel, horizontal, and predictable.

Figure 6.5 expresses the relationship between various stopping rules and MCMT performance when using different item banks and conditioning on the compensatory or non-compensatory classification bound function. As shown in Figure 6.5, one finds scant evidence of an interaction between item bank and stopping rule when assuming a compensatory classification bound. All of the pink points are above and to the left of all of the blue points in the upper left panel of Figure 6.5. Moreover, the order of stopping rules within each item bank on the accuracy and test length using a compensatory classification bound function is nearly identical for both the within-item and between-item multidimensional banks, as shown in the upper right panel of Figure 6.5. Therefore,

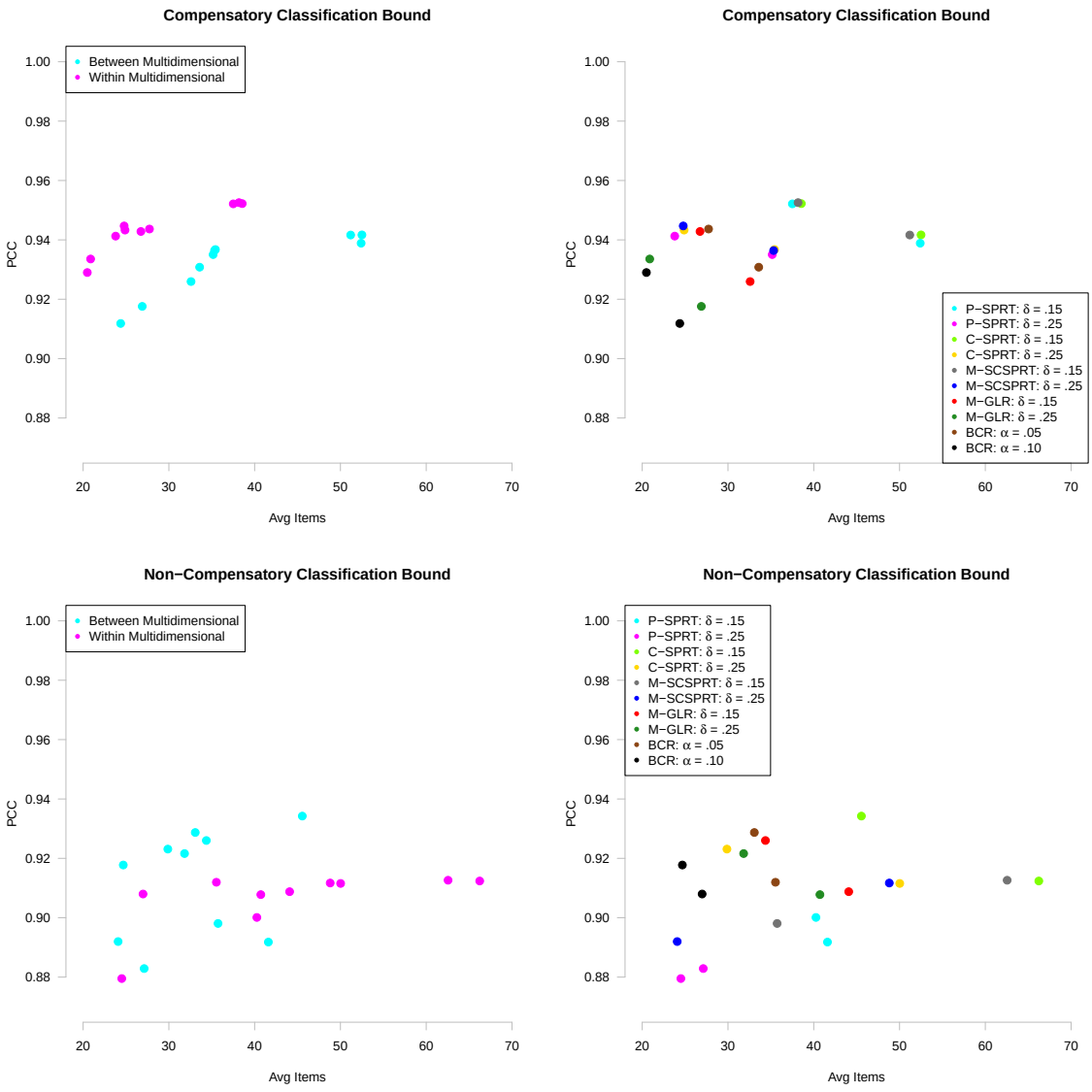

Figure 6.5: Scatterplots of the percent classified correctly (PCC) by average number of items administered based on the interaction between item bank and stopping rule using either a compensatory classification bound function (top panels) or a non-compensatory classification bound function (bottom panels). The left panels are color coded according to item bank, whereas the right panels are color coded according to stopping rule.

every stopping rule performs better, on average (i.e., has a greater overall accuracy and a shorter average test length), when using the within-item multidimensional bank, and the gain in accuracy and test length appears to be similar regardless of stopping rule. This relationship between item bank, stopping rule, and test statistics is only approximately true when examining the non-compensatory classification bound function, as evidenced by the lower panels of Figure 6.5. Notice that most of the light blue points are above and to the left of the pink points. In fact, as in the compensatory classification bound, the stopping rules represented by those six blue points have better classification accuracy as well as shorter tests when using the between-item multidimensional bank. In the lower right panel of Figure 6.5, the black, yellow, dark green, brown, red, and light green points are above and to the left when using the between-item multidimensional bank as compared to the within-item multidimensional bank. The only outliers to this general trend are points representing the P-SPRT and M-SCSPRT conditions, those conditions already identified as poorly performing when using the non-compensatory classification bound function. The loss conception of performance cleanly expresses this general trend, as shown on the lower panels of Figure 6.4. When using the compensatory classification bound, the loss trend shape is nearly identical for a given stopping rule if selecting items from either item bank. However, the loss curves are all lower throughout the entire range of P when selecting items from the within-item bank as compared to the between-item bank. One finds a similar relationship when examining loss for the non-compensatory classification bound with a few exceptions. First, the between-item bank results in the lowest loss when using the non-compensatory classification bound function. Second, P-SPRT and M-SCSPRT result in very large loss regardless of item bank. Finally, all of the other stopping rules yield similar relative loss after $P \approx 1{,}000$.

To supplement conclusions drawn from Figures 6.1–6.5, several ANOVA tables were constructed, four of which will be presented forthwith: (1) Tables 6.1 and 6.3 indicate

the effect of various factors on mean test length for either the compensatory or non-compensatory classification bound functions, respectively; and (2) Tables 6.2 and 6.4 summarize the effect of various factors on classification accuracy. I have also included tables describing ANOVAs when predicting three values of mean loss from the same factors in Appendix B. As I suggested in my earlier study on unidimensional classification testing, despite ANOVA being potentially inappropriate given the research design, several authors have used ANOVA to obtain descriptive measures of variance accounted for by factors in a Monte Carlo study (e.g., Guyer & Weiss, 2009). All of the following tables present both $\eta^2 = \frac{\text{SSF}}{\text{SST}}$ and $\omega^2 = \frac{\text{SSF}-\text{df}_F \times \text{MSE}}{\text{SST}+\text{MSE}}$, where SSF is the sums of squares for factor F, and SST is the total sums of squares. Note that η^2 is a positively biased (similar to R^2) estimate of the proportion of variance accounted for by each factor, and ω^2 is a less biased estimated obtained by rearranging formulas for the expected mean squares (e.g., Abelson, 1985; Olejnik & Algina, 2000). Also notice that the η^2 values and ω^2 values are nearly identical across all of the tables, so I will only describe η^2.

Based on Tables 6.1 and 6.2, stopping rule, item bank, the true correlation between latent ability, and the interaction between item bank and stopping rule account for most of the variability in both test length and classification accuracy when using a compensatory classification bound function. As expected, stopping rule accounts for most of the observable variance in both test length ($\eta^2 = .676$) and classification accuracy ($\eta^2 = .466$) with item bank accounting for most of the remaining variance ($\eta^2 = .230$ and $\eta^2 = .233$ in either case). Therefore, choosing the inappropriate stopping rule or item bank will have the greatest effect on MCMT properties. Moreover, one also finds very little interaction between ability correlation and the other factors. In fact, the two-way and three-way interactions of correlation and any of the other factors result in the smallest effect sizes in either of the tables. Interestingly, one finds a small but noticeable effect of the interaction between bank and stopping rule on both test length

Table 6.1: The sums of squares (Sum Sq.), $\eta^2 = \frac{\text{SSF}}{\text{SST}}$, and $\omega^2 = \frac{\text{SSF} - \text{df}_F \times \text{MSE}}{\text{SST} + \text{MSE}}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting mean test length given a compensatory classification bound function. The ANOVA was run with all main effects, two-way interactions, and three-way interactions.

Variance Type	Sum Sq.	η^2	ω^2
Correlation (Cor)	947.55	.032	.032
Item Bank (Bank)	6901.34	.230	.230
Select Alg. (Select)	382.07	.013	.013
Stop Rule (Stop)	20301.92	.676	.676
Cor by Bank	20.47	.001	.001
Cor by Select	6.07	.000	.000
Cor by Stop	49.39	.002	.002
Bank by Select	214.77	.007	.007
Bank by Stop	1030.04	.034	.034
Select by Stop	84.24	.003	.003
Cor by Bank by Sel	10.82	.000	.000
Cor by Bank by Stop	1.21	.000	.000
Cor by Sel by Stop	4.52	.000	.000
Bank by Sel by Stop	40.64	.001	.001
Residuals	16.31		
Total	30011.36		

($\eta^2 = .034$) and classification accuracy ($\eta^2 = .025$) despite minimal visual evidence of this effect in Figure 6.5. All of the other conditions result in a smaller-than-noticeable effect and thus are not worth further discussion. One can see the very small relationship between item selection and stopping rule on the accuracy and average test length of MCMTs when using a compensatory classification bound function in Figure C.8 of Appendix C.

With respect to the non-compensatory classification bound function, one finds similar effects of stopping rule ($\eta^2 = .488$), item bank ($\eta^2 = .198$) and the interaction

Table 6.2: The sums of squares (Sum Sq.), $\eta^2 = \frac{\text{SSF}}{\text{SST}}$, and $\omega^2 = \frac{\text{SSF} - \text{df}_F \times \text{MSE}}{\text{SST} + \text{MSE}}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting mean classification accuracy given a compensatory classification bound function. The ANOVA was run with all main effects, two-way interactions, and three-way interactions.

Variance Type	Sum Sq.	η^2	ω^2
Correlation (Cor)	0.00841	.185	.185
Item Bank (Bank)	0.01055	.233	.232
Select Alg. (Select)	0.00054	.012	.011
Stop Rule (Stop)	0.02115	.466	.464
Cor by Bank	0.00001	.000	.000
Cor by Select	0.00004	.001	.000
Cor by Stop	0.00010	.002	.000
Bank by Select	0.00044	.010	.009
Bank by Stop	0.00111	.025	.023
Select by Stop	0.00066	.015	.007
Cor by Bank by Select	0.00006	.001	.001
Cor by Bank by Stop	0.00014	.003	.001
Cor by Select by Stop	0.00038	.008	.004
Bank by Sel by Stop	0.00053	.012	.000
Residuals	0.00123		
Total	0.04535		

between item bank and stopping rule ($\eta^2 = .175$) on average test length, as shown in Table 6.3. Therefore, similar factors result in differences in test lengths for both the compensatory and non-compensatory classification bound functions. In fact, the decreased effect of item bank on test length when using the non-compensatory classification bound is probably due to the four poor performing conditions (i.e., both of the P-SPRT and M-SCSPRT conditions) that perform differently than the rest of the stopping rules when adopting a non-compensatory classification bound function. Notice how those are the only four conditions that lead to conflicting trends in the loss lines of Figure 6.4. When examining accuracy using the non-compensatory classification bound function, one finds the strongest effects for stopping rule ($\eta^2 = .285$), item selection by stopping rule ($\eta^2 = .229$), and the three way interaction between item bank, item selection algorithm, and stopping rule ($\eta^2 = .188$). Not surprisingly, these relationships are entirely due to the P-SPRT and M-SCSPRT stopping rules. If running an ANOVA after eliminating both P-SPRT and M-SCSPRT stopping rules, then item bank has the strongest effect on classification accuracy ($\eta^2 = .507$) followed by item selection ($\eta^2 = .108$) and stopping rule ($\eta^2 = .104$). The other factors and interactions have much smaller effects on classification accuracy.

6.2 Results 2: Conditional on Specific Ability Vectors

This section section examines the accuracy and test length of various stopping rules and item selection algorithms conditional on particular ability vectors[3]. Most mastery tests need not classify all examinees with equal precision. Examinees with ability vectors close to the classification bound should require more items to determine the appropriate

[3]All of the figures generated by using statistics conditional on particular ability vectors are presented in Appendix D.

Table 6.3: The sums of squares (Sum Sq.), $\eta^2 = \frac{\text{SSF}}{\text{SST}}$, and $\omega^2 = \frac{\text{SSF} - \text{df}_F \times \text{MSE}}{\text{SST} + \text{MSE}}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting mean test length given a non-compensatory classification bound function. The ANOVA was run with all main effects, two-way interactions, and three-way interactions.

Variance Type	Sum Sq.	η^2	ω^2
Correlation (Cor)	23.52	.000	.000
Item Bank (Bank)	9372.46	.198	.198
Select Alg. (Select)	2499.74	.053	.053
Stop Rule (Stop)	23168.56	.488	.488
Cor by Bank	31.34	.001	.001
Cor by Select	11.53	.000	.000
Cor by Stop	44.09	.001	.001
Bank by Select	529.32	.011	.011
Bank by Stop	8308.66	.175	.175
Select by Stop	2147.33	.045	.045
Cor by Bank by Select	3.43	.000	.000
Cor by Bank by Stop	20.03	.000	.000
Cor by Select by Stop	45.52	.001	.001
Bank by Select by Stop	1166.62	.025	.024
Residuals	63.67		
Total	47435.8		

Table 6.4: The sums of squares (Sum Sq.), $\eta^2 = \frac{\text{SSF}}{\text{SST}}$, and $\omega^2 = \frac{\text{SSF} - \text{df}_F \times \text{MSE}}{\text{SST} + \text{MSE}}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting mean classification accuracy given a non-compensatory classification bound function. The ANOVA was run with all main effects, two-way interactions, and three-way interactions.

Variance Type	Sum Sq.	η^2	ω^2
Correlation (Cor)	.00228	.014	.014
Item Bank (Bank)	.00202	.013	.012
Select Alg. (Select)	.01002	.062	.062
Stop Rule (Stop)	.04586	.285	.283
Cor by Bank	.00033	.002	.002
Cor by Select	.00052	.003	.003
Cor by Stop	.00151	.009	.008
Bank by Select	.00922	.057	.057
Bank by Stop	.01417	.088	.087
Select by Stop	.03695	.229	.224
Cor by Bank by Select	.00008	.000	.000
Cor by Bank by Stop	.00227	.014	.013
Cor by Select by Stop	.00207	.013	.007
Bank by Select by Stop	.03036	.188	.183
Residuals	.00343		
Total	.16110		

classification. Moreover, classification near the bound separating master from non-master will necessarily be less precise than classification decisions for examinees well within any category.

The results of the previous section influenced those conditions that I chose to examine in more depth. For instance, because latent trait correlation did not appear to have much of an effect on test length or classification accuracy (and next to no interaction with any of the other factors), I decided to only examine the moderate correlation of $\rho = .33$. Moreover, I chose to only use the within-item multidimensional bank for the compensatory classification bound function and the between-item multidimensional bank for the non-compensatory classification bound function due to their overall performance when simulating from a distribution. Finally, I decided to eliminate the D-FI item selection algorithm due to inefficiency and the P-SPRT stopping rule due to poor performance when using a non-compensatory classification bound function. After eliminating those conditions deemed inefficient, inaccurate, or uninteresting, I retained 1 ability correlation ($\rho = .33$) $\times$ 1 classification bound function (within-item multidimensionality for the compensatory classification bound or between-item multidimensionality for the non-compensatory classification bound) $\times$ 4 item selection algorithms (L-FI, L-ELR, L-KL, and S-KL) $\times$ 6 stopping rules (C-SPRT and M-SCSPRT with $\delta = .25$, M-GLR with $\delta \in \{.15, .25\}$, and BCR with $\alpha \in \{.05, .10\}$) $\times$ 2 classification bound functions = 48 total conditions.

To determine the conditional classification accuracy and test length, 64 ability vectors were chosen to be equidistant along a 8×8 square with $\theta_{ik} \in \{-.7, -.5, -.3, -.1, +.1, +.3, +.5, +.7\}$, rotated based on the correlation matrix with $\rho = .33$, and then tested 1,000 times according to each combination of conditions.

Figure 6.6 displays the conditional accuracy rate for several ability vectors and several item selection algorithms given a compensatory classification bound function with

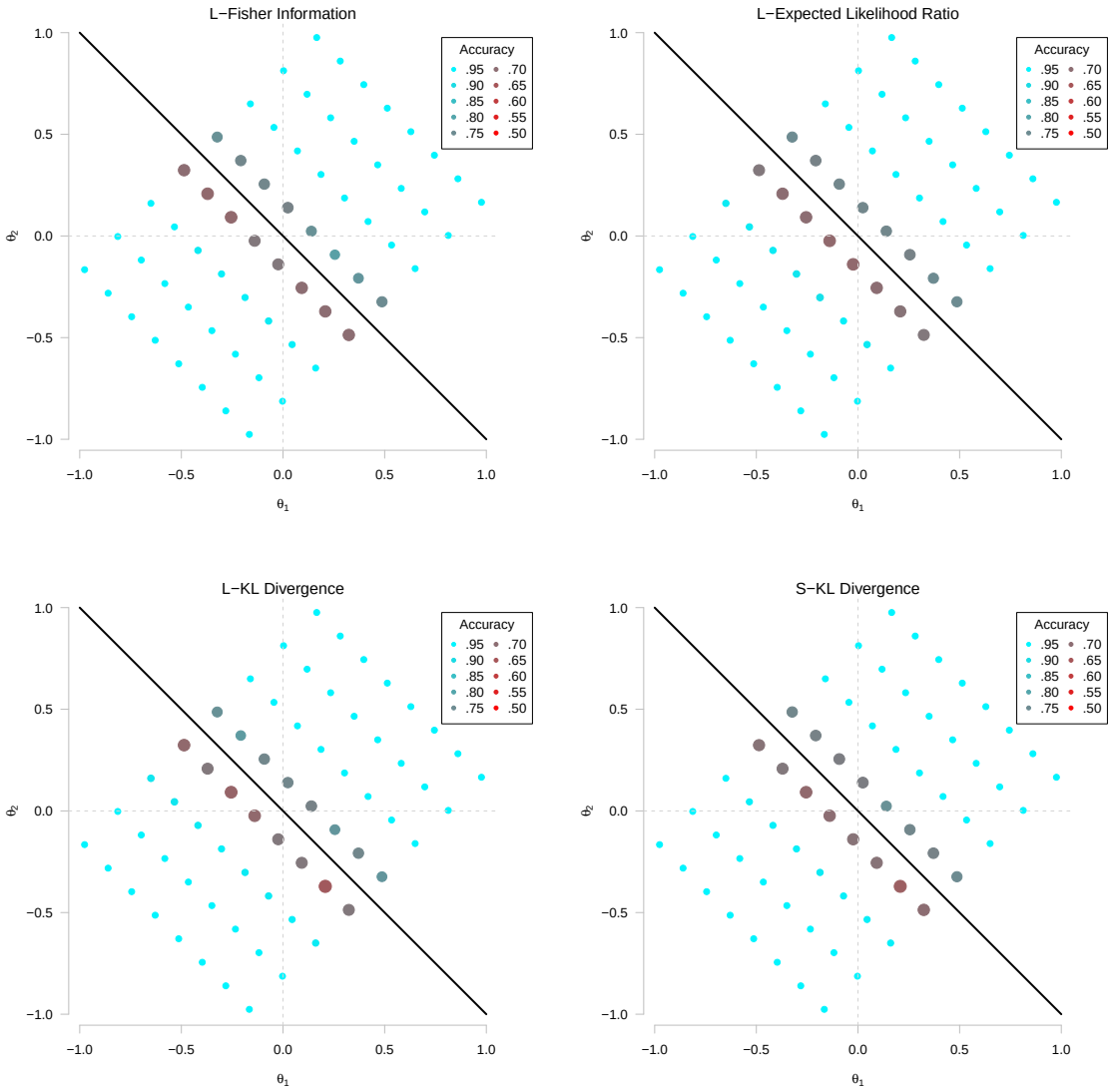

Figure 6.6: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the compensatory classification bound function and the C-SPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

$\rho = .33$, a within-item multidimensional bank, and the C-SPRT stopping rule. For all of the figures in this section, the size and color of the bubbles represents the accuracy rate, test length, or loss function: (1) Red large bubbles implies poor accuracy, long tests, or large loss; (2) blue-green medium bubbles implies medium accuracy, moderately long tests, or medium-large loss; and (3) light-blue small bubbles implies good accuracy, short tests, or small loss. One thing to spot across all of the accuracy plots in this section is their similar appearance. Notice that when using the C-SPRT stopping rule with a compensatory classification bound function, all of the item selection algorithms result in very similar accuracy rates for all values of $\boldsymbol{\theta}$. For these conditions, examinees slightly below the classification bound have poorer classification accuracy than examinees slightly above the classification bound, as evidenced by the slightly larger and redder bubbles below the classification bound function than above the classification bound function. But ability vectors slightly further away from the classification bound function are all classified with the same, exceptional accuracy rate. For the compensatory classification bound function, only M-GLR with $\delta = .25$ and BCR with $\alpha = .10$ result in differently shaped plots, as shown in Figures 6.7 and 6.8. Yet the only difference between Figures 6.6 and 6.7 or 6.8 is that when using the latter stopping rules, the first two rows of points below the classification bound are slightly larger and redder.

As in the unidimensional case (see Nydick, 2012), one finds larger discrepancies between conditions when examining test length rather than classification accuracy. For example, compare the general pattern of results using a C-SPRT stopping rule with $\delta = .25$ to that using a M-GLR stopping rule with $\delta = .15$ or a BCR stopping rule with $\alpha = .05$, as shown in Figures 6.9–6.11. The particular C-SPRT chosen yields much shorter tests for true ability vectors close to the classification bound as compared to the M-GLR and BCR conditions. For these conditions, both the first and second lines of points are larger when using M-GLR and BCR as compared to C-SPRT. But unlike

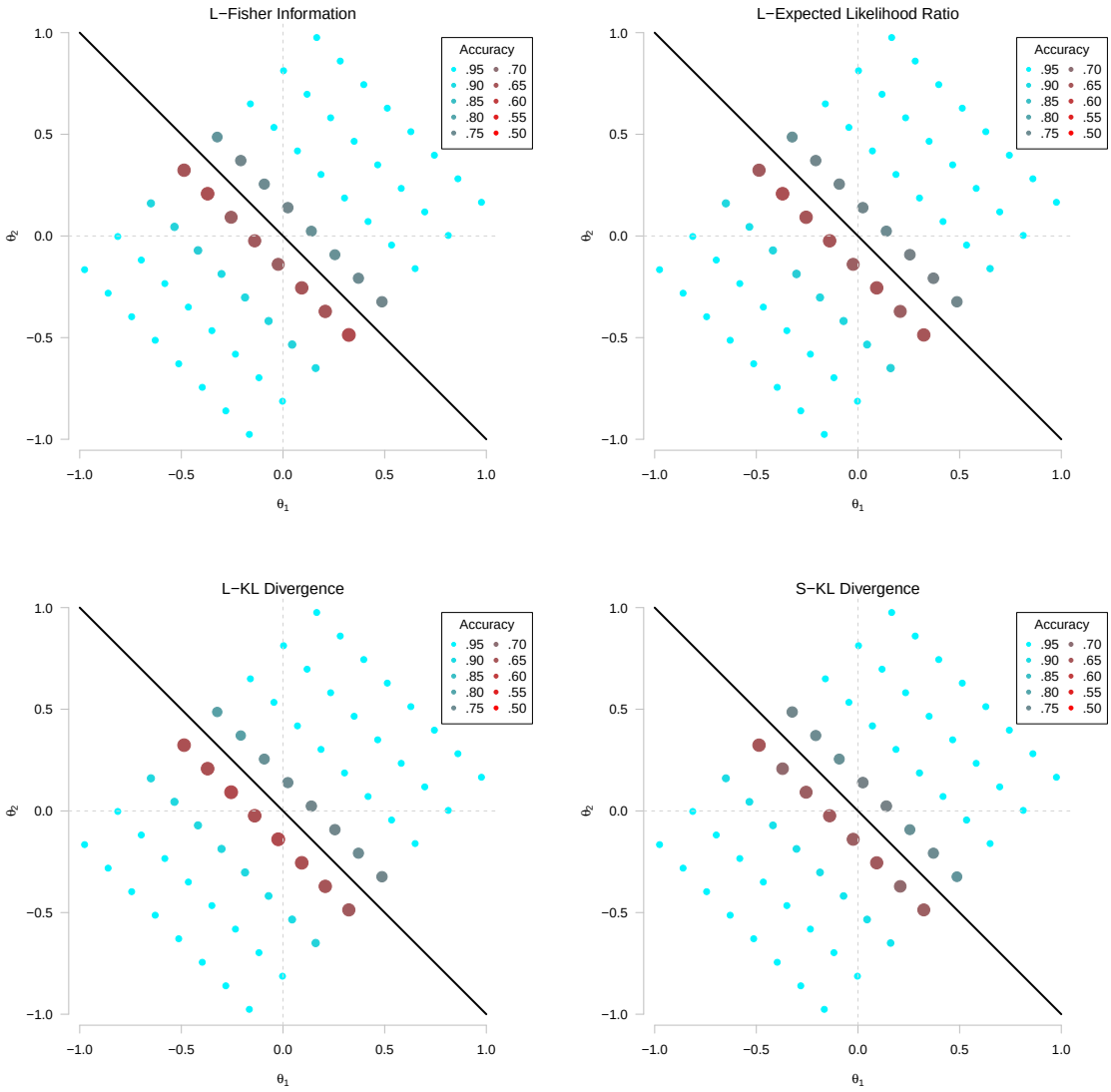

Figure 6.7: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the compensatory classification bound function and the M-GLR stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

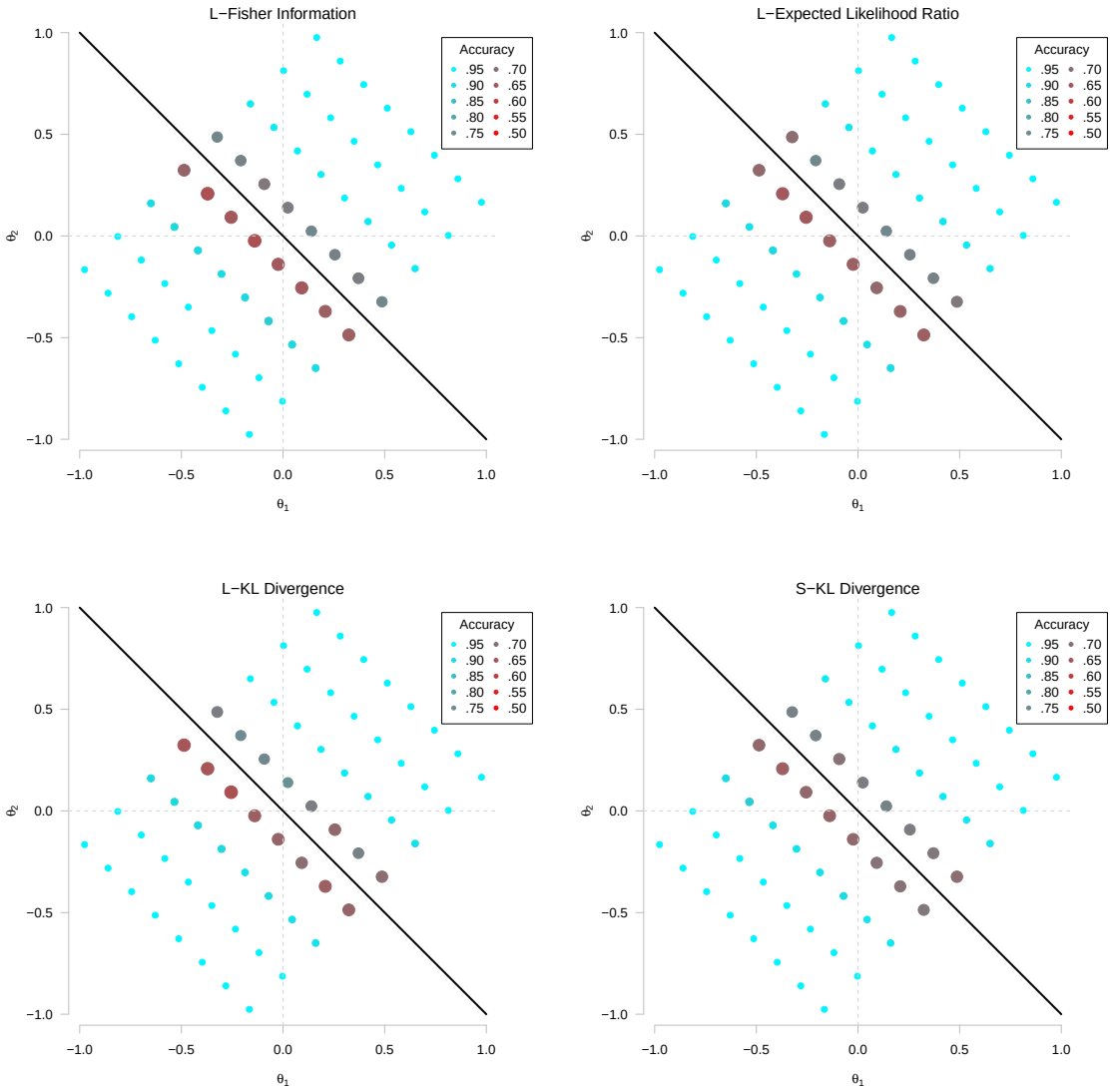

Figure 6.8: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the compensatory classification bound function and the BCR stopping rule with $\alpha = .10$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

classification accuracy, the change in test length is equivalent both above and below the classification bound function for all stopping rules. Yet when using M-GLR with $\delta = .25$ or BCR with $\alpha = .10$, test length is shorter for true ability vectors close to the classification bound function than C-SPRT with $\delta = .25$ (see, for example, Figure D.17 or D.19 in Appendix D). This finding emphasizes the important point that when using a compensatory classification bound function, parameters of a given stopping rule matter more than the specific stopping rule chosen, and test length is more affected by these parameter values than accuracy rates. Moreover, distance from the compensatory classification bound function is more relevant than location, at least for the points within the chosen rectangle. Notice that the changes in test length or accuracy given different stopping rules affect all points equally as long as those points are on the same line parallel to the $\theta_1 + \theta_2 = 0$ classification bound function. One also finds scant evidence of differences between different item selection algorithms for the compensatory classification bound function given a particular stopping rule. All quadrants within any of the presented figures are similar in shape, color, and pattern. These results parallel and reinforce those results presented in the previous section by implying that the small differences in accuracy and test length for various item selection algorithms when using the compensatory classification bound function are the same regardless of true ability.

One finds similar trends for both conditional accuracy and test length when using a non-compensatory classification bound function, as shown in Figures 6.12–6.14. I chose to present C-SPRT with $\delta = .25$, M-GLR with $\delta = .25$, and BCR with $\alpha = .10$ due to their relative dissimilarities. As in the compensatory classification bound, one finds slightly larger bubbles a similar distance below the non-compensatory classification bound function than above the non-compensatory classification bound function. Therefore, in almost all cases, simulees were classified better for true ability vectors above the classification bound function. This general trend does not hold for the middle two

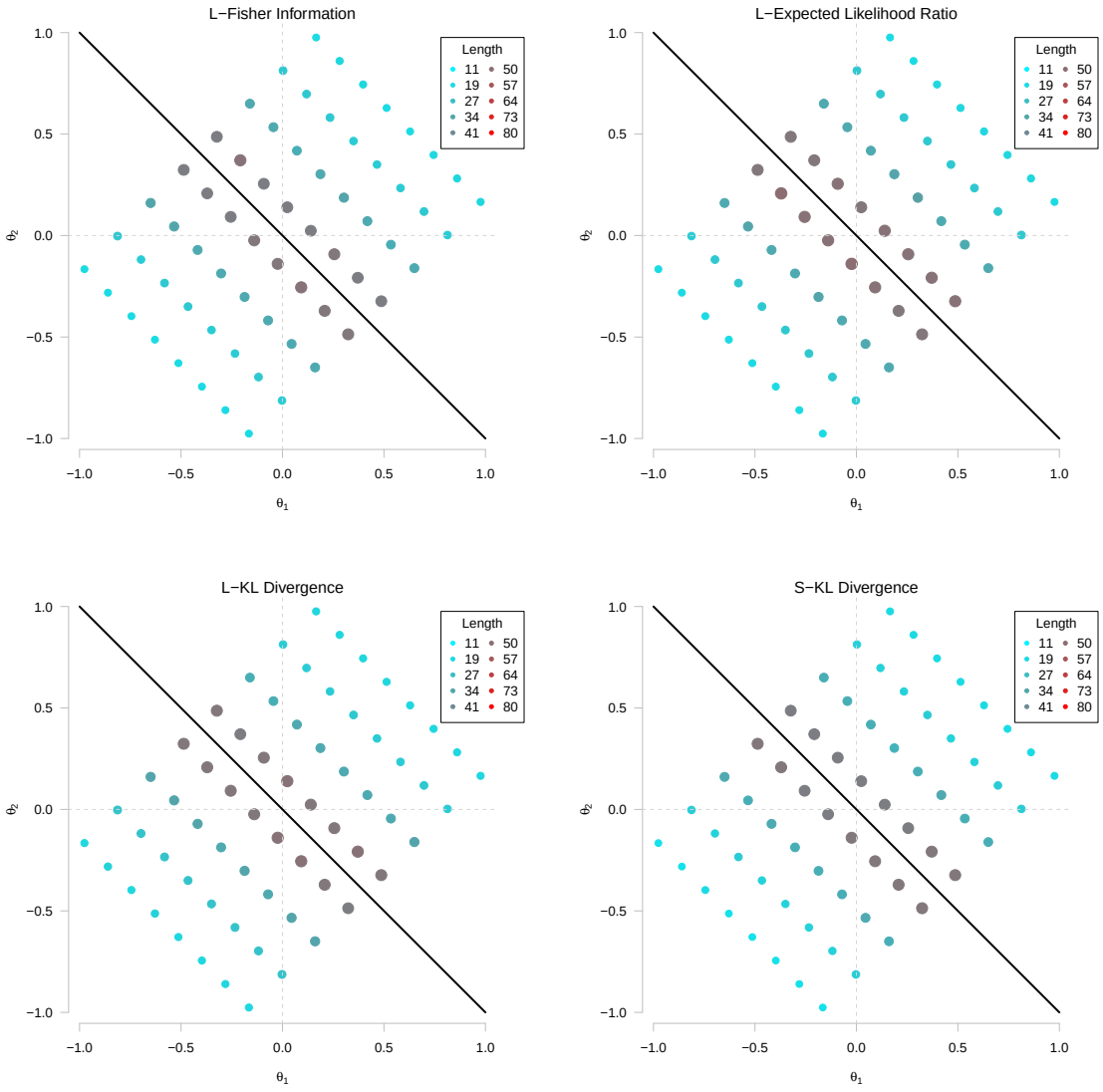

Figure 6.9: Scatterplots of the conditional average test length for various vectors of true ability when using the compensatory classification bound function and the C-SPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

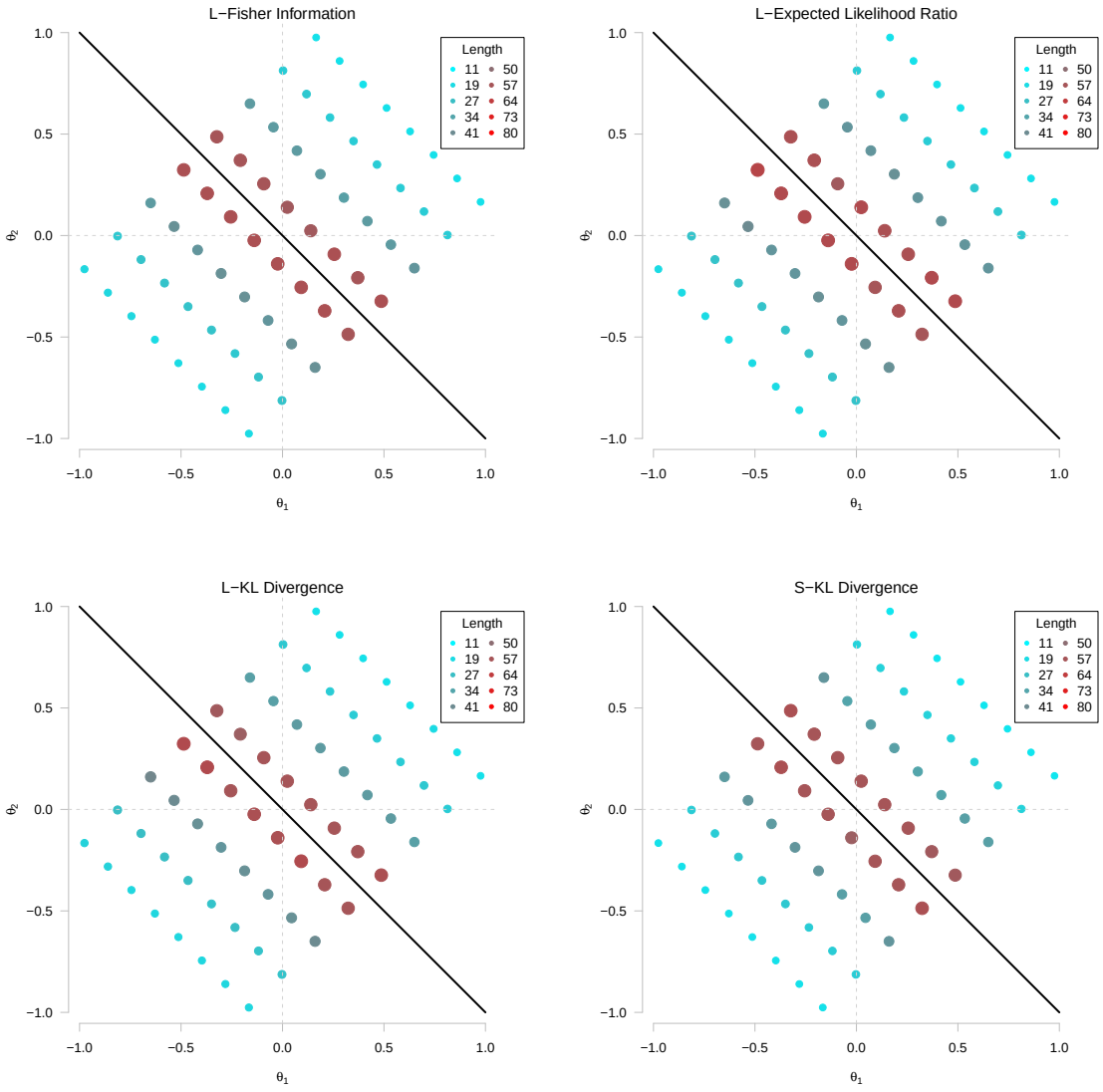

Figure 6.10: Scatterplots of the conditional average test length for various vectors of true ability when using the compensatory classification bound function and the M-GLR stopping rule with $\delta = .15$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

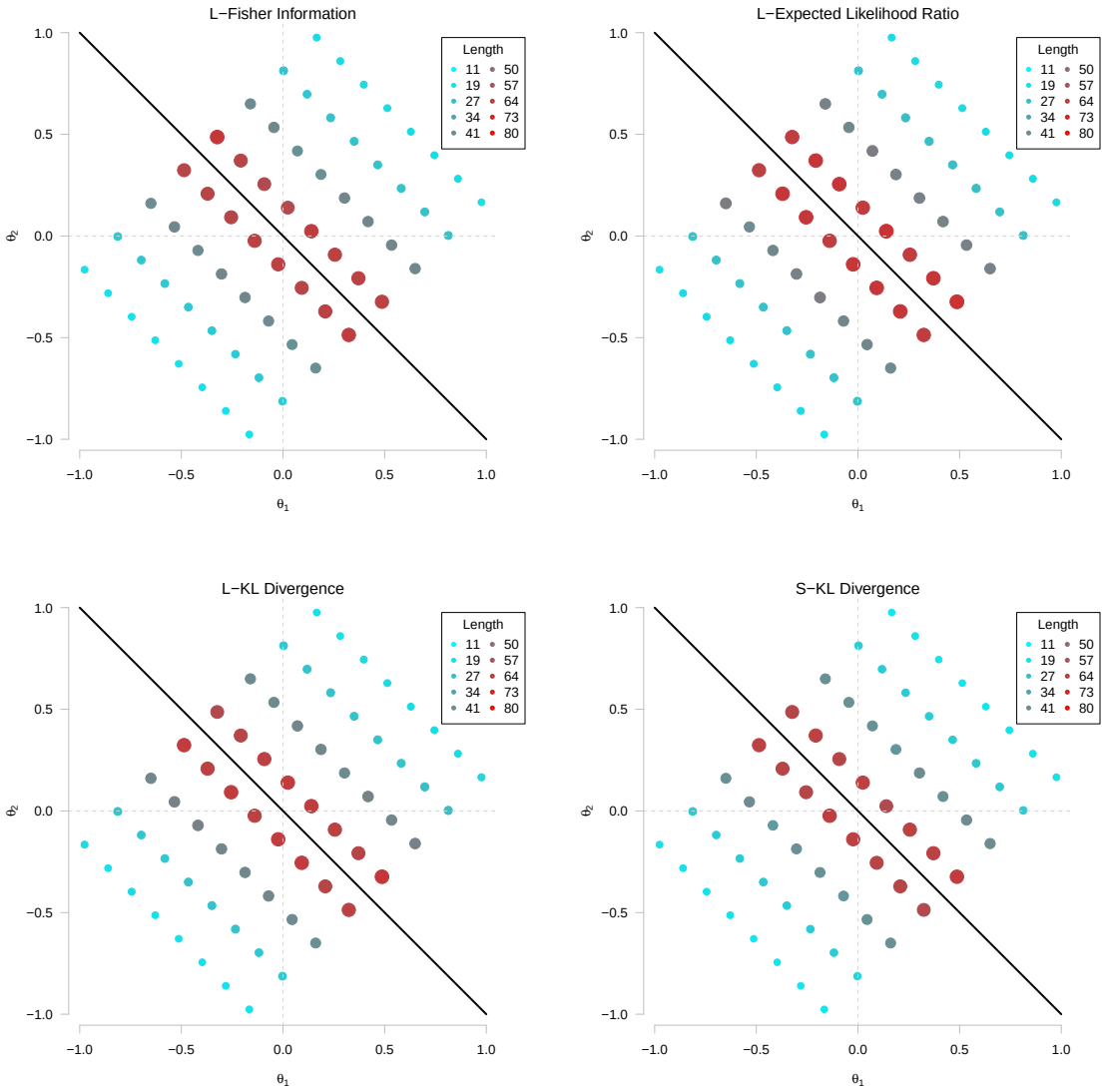

Figure 6.11: Scatterplots of the conditional average test length for various vectors of true ability when using the compensatory classification bound function and the BCR stopping rule with $\alpha = .05$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

dotted lines along the longer side of the rectangle. For those ability vectors near the classification bend, simulees were more accurately classified when their true ability lie below the classification bound function rather than above. The divergent trend when classifying simulees with true ability near the classification bend is most noticeable when using the BCR stopping rule with $\alpha = .10$, as shown in Figure 6.14. In fact, BCR with $\alpha = .10$ had similar accuracy rates to the alternative stopping rules for all true ability vectors below the classification bound but much worse accuracy rates for ability vectors above the classification bound.

A notable exception to the general pattern of accuracy results when using the non-compensatory classification bound function is the M-SCSPRT stopping rule, as shown in Figure 6.15. Recall that when using the non-compensatory classification bound, the M-SCSPRT stopping rule and L-ELR item selection algorithm resulted in accuracy rates well below most other conditions. Based on the upper-left panel of Figure 6.15, one finds that the classification accuracy is much worse for practically all latent ability vectors when using the M-SCSPRT stopping rule with the L-ELR item selection algorithm than any other combination of conditions. In fact, if examining M-SCSPRT, one finds similar trends to C-SPRT (or M-GLR/BCR) for all except the L-ELR condition. A probable reason for the poor classification accuracy when selecting items via L-ELR and terminating a test by means of M-SCSPRT is that M-SCSPRT requires an estimate of future items to determine classification probabilities, and L-ELR poorly chooses those future items given inaccurate $\hat{\boldsymbol{\theta}}_i$s early in a test.

Figures 6.16–6.18 display the conditional, average test length corresponding to those conditions presented in Figures 6.12–6.14. Unlike the trends in test length when using a compensatory classification bound, both the item selection and the stopping rule have distinct effects on the overall average test length as well as the shift in short/long test length regions on a given plot. For instance, consider Figure 6.16, which depicts the

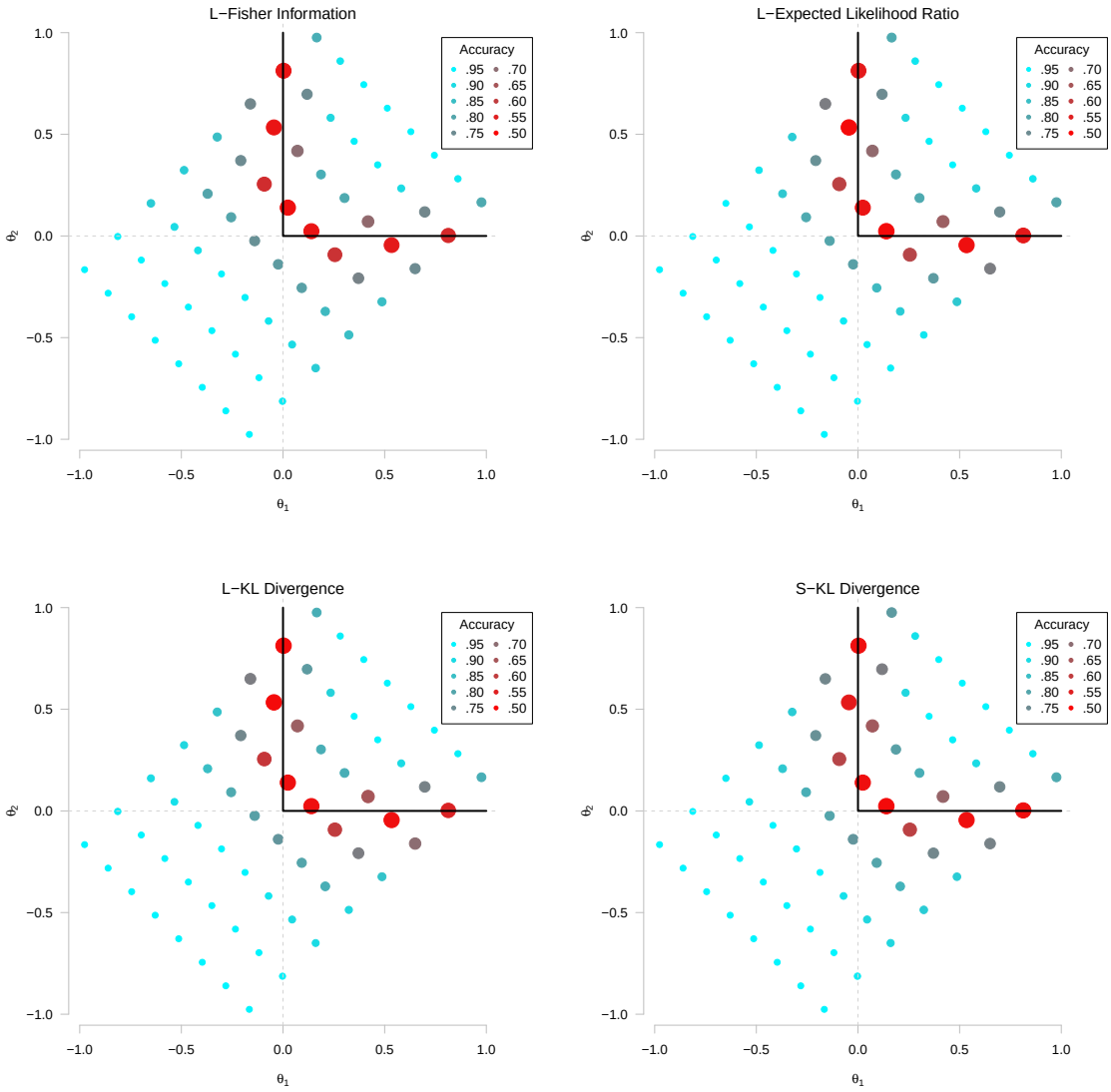

Figure 6.12: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the non-compensatory classification bound function and the C-SPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

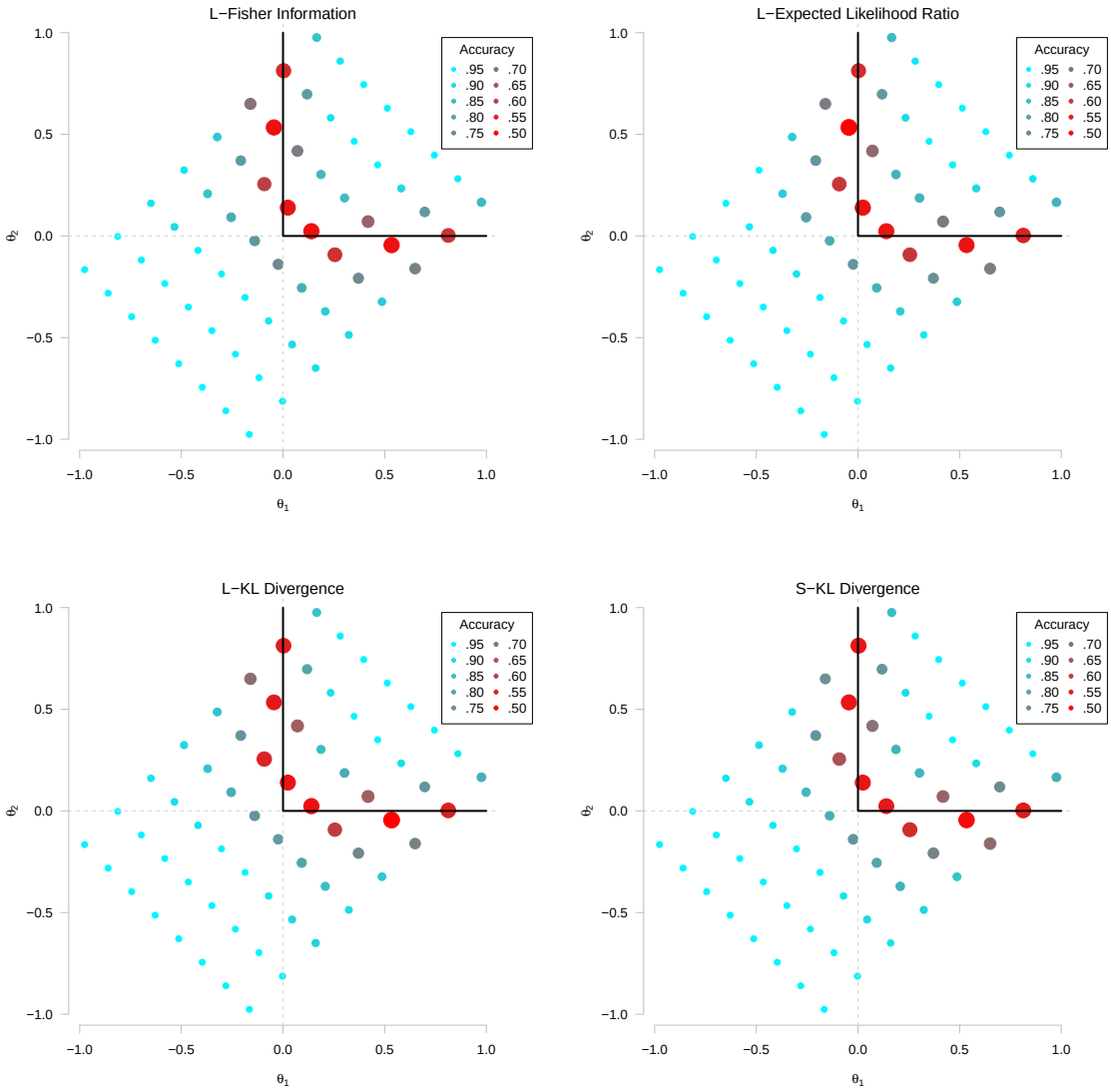

Figure 6.13: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the non-compensatory classification bound function and the M-GLR stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

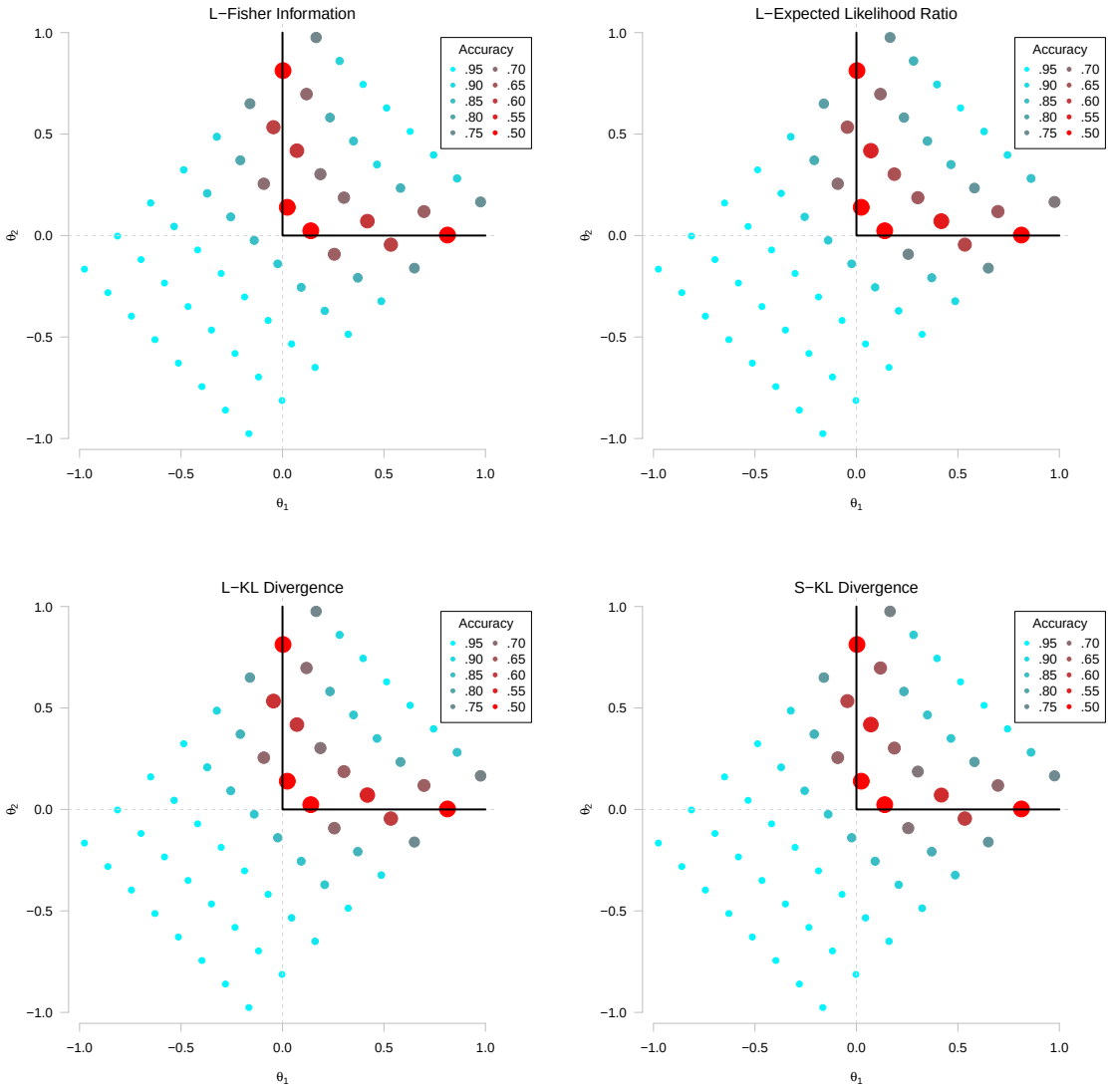

Figure 6.14: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the non-compensatory classification bound function and the BCR stopping rule with $\alpha = .10$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

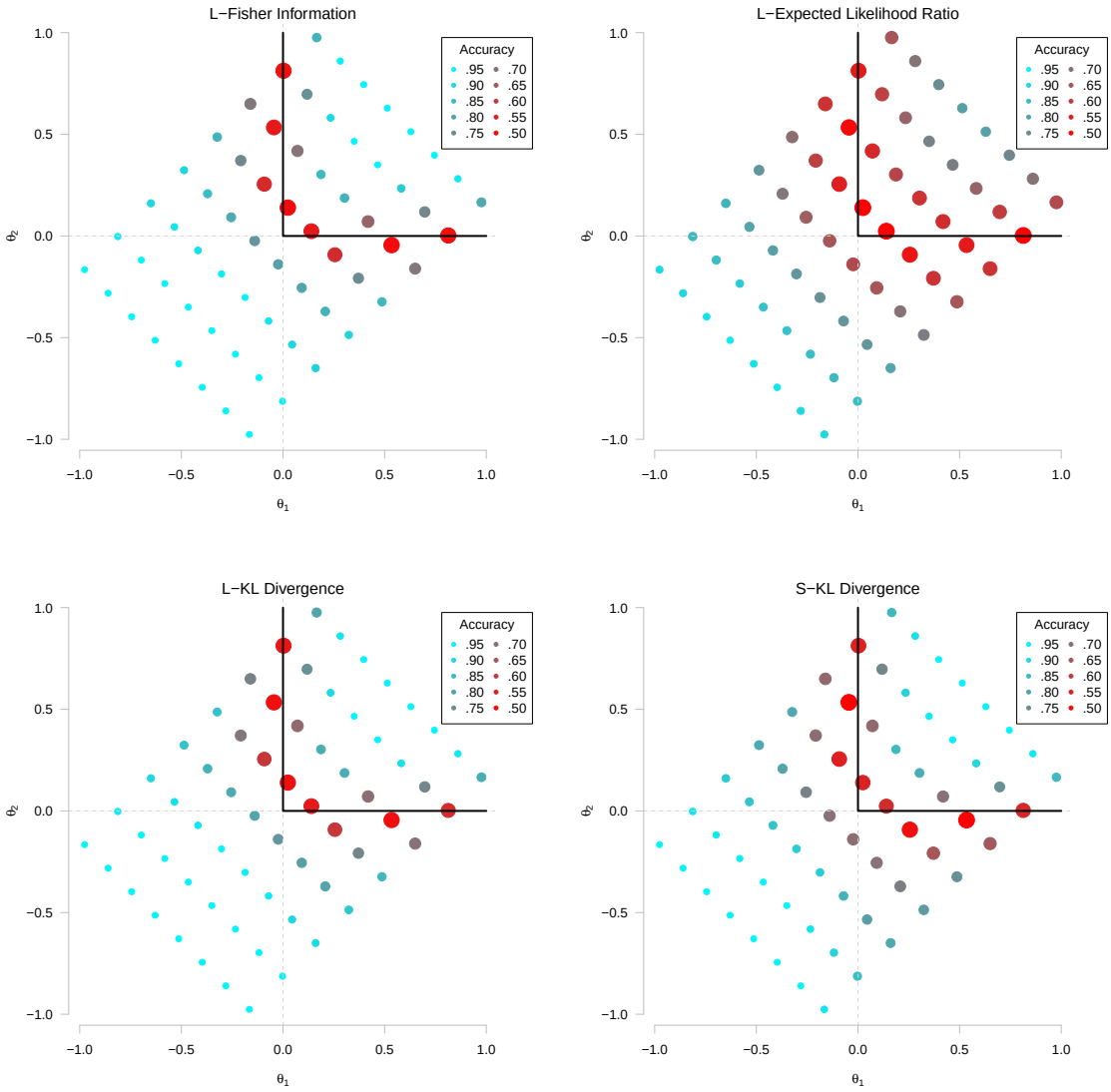

Figure 6.15: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the non-compensatory classification bound function and the M-SCSPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

average test length for four item selection algorithms when implementing the C-SPRT stopping rule. Given C-SPRT with $\delta = .25$, all four item selection algorithms yield longer tests above the classification bound function than below the classification bound function. The trend across all four C-SPRT test length plots is similar to the accuracy trend with the exception of a lesser gradation in test length when increasing one's distance from the classification bound. However, as is clearly shown, L-ELR results in longer tests for all true ability vectors than S-KL, which results in longer tests for all true ability vectors than L-FI or L-KL. Note that the M-GLR stopping rule with $\delta = .25$ evinces the same pattern across item selection algorithms as C-SPRT with $\delta = .25$, as shown in Figure 6.17. BCR with $\alpha = .10$ exhibits much smaller differences in test length than either M-GLR or C-SPRT, as shown by Figure 6.18. Yet the major difference between BCR and the other stopping rules is the location of efficient performance. Note that BCR with $\alpha = .10$ yields moderately long tests for true ability vectors above the classification bound and much shorter tests for ability vectors below the classification bound. The center of inefficiency is thus located well within the upper classification category. However C-SPRT with M-GLR, both with $\delta = .25$, yield tests with the center of inefficiency closer to the classification bound function. The differences between M-GLR and BCR are more apparent when comparing M-GLR with $\delta = .25$ (Figure 6.17) to the less efficient BCR with $\alpha = .05$ (Figure 6.19). With respect to the latter stopping rule, one can clearly see a cluster of large red points well within the upper classification category, and with respect to the former stopping rule, one can see the cluster of large red points shifted across the classification bound. Therefore, if test practitioners want to ensure that true masters take the longest tests, then the practitioner should use the BCR stopping rule; however, if these practitioners only want to ensure that examinees with true ability within the indifference region (on either side of the classification bound function) take the longest tests, then s/he should use C-SPRT or M-GLR.

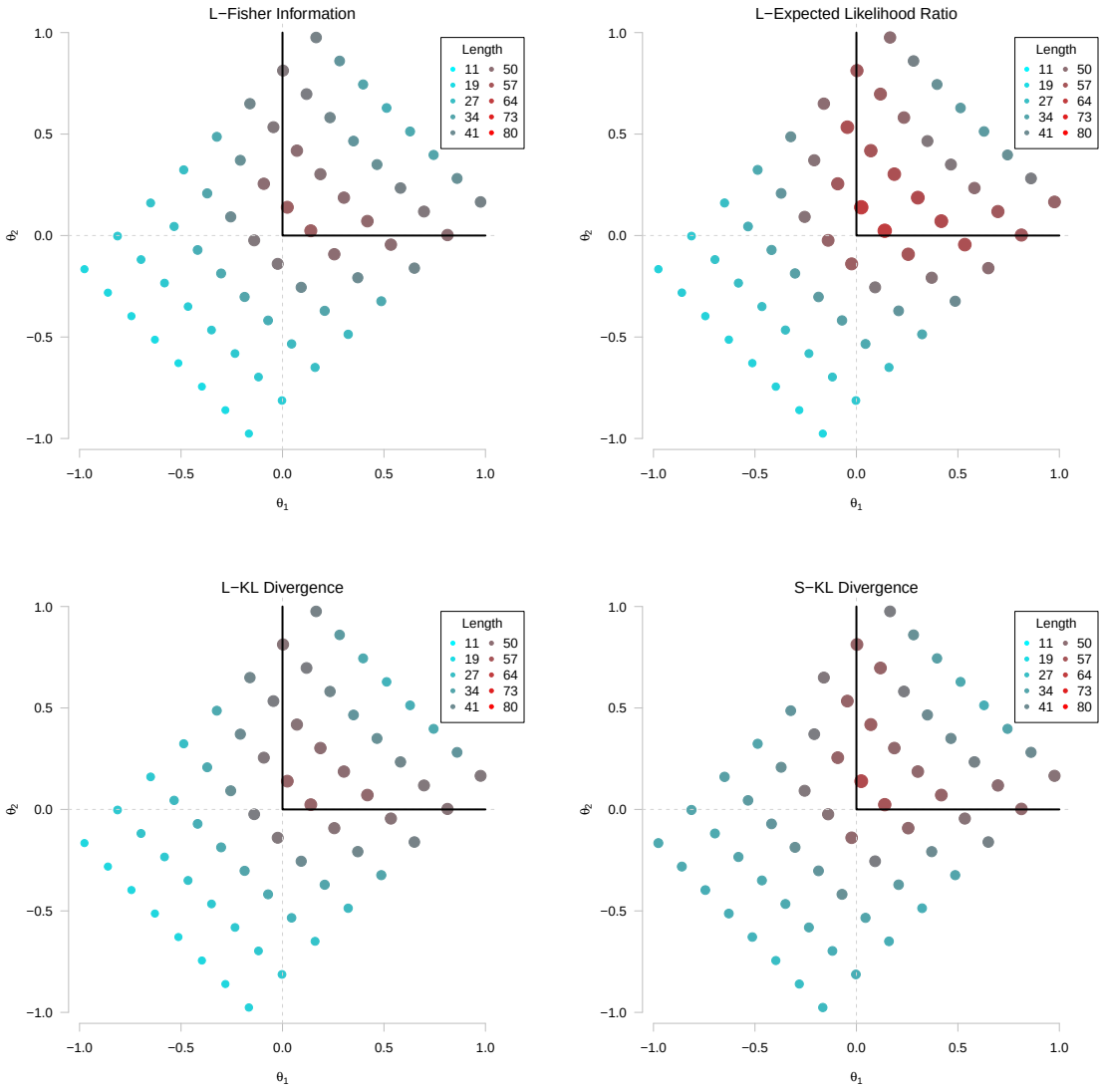

Figure 6.16: Scatterplots of the conditional average test length for various vectors of true ability when using the non-compensatory classification bound function and the C-SPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

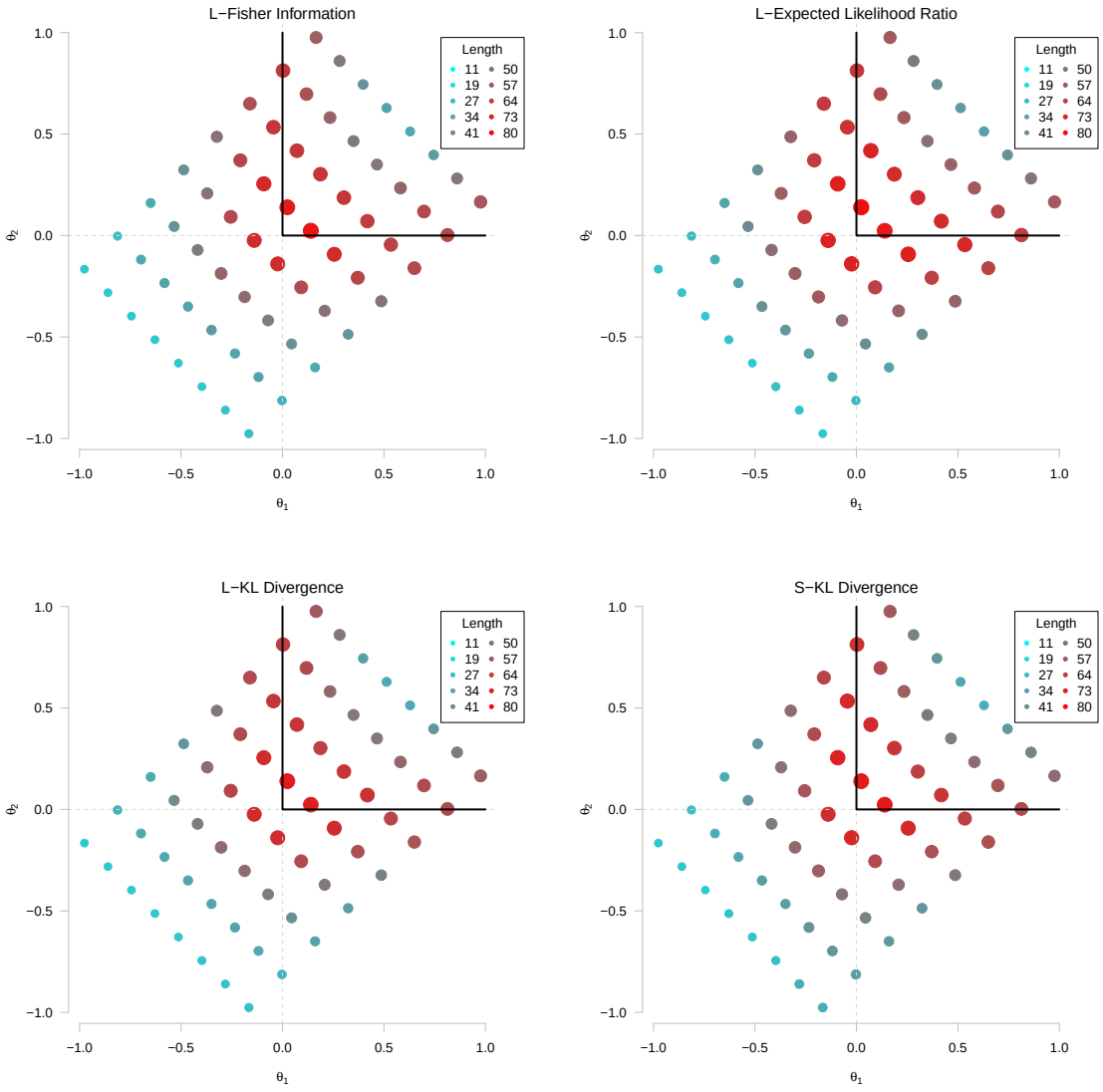

Figure 6.17: Scatterplots of the conditional average test length for various vectors of true ability when using the non-compensatory classification bound function and the M-GLR stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

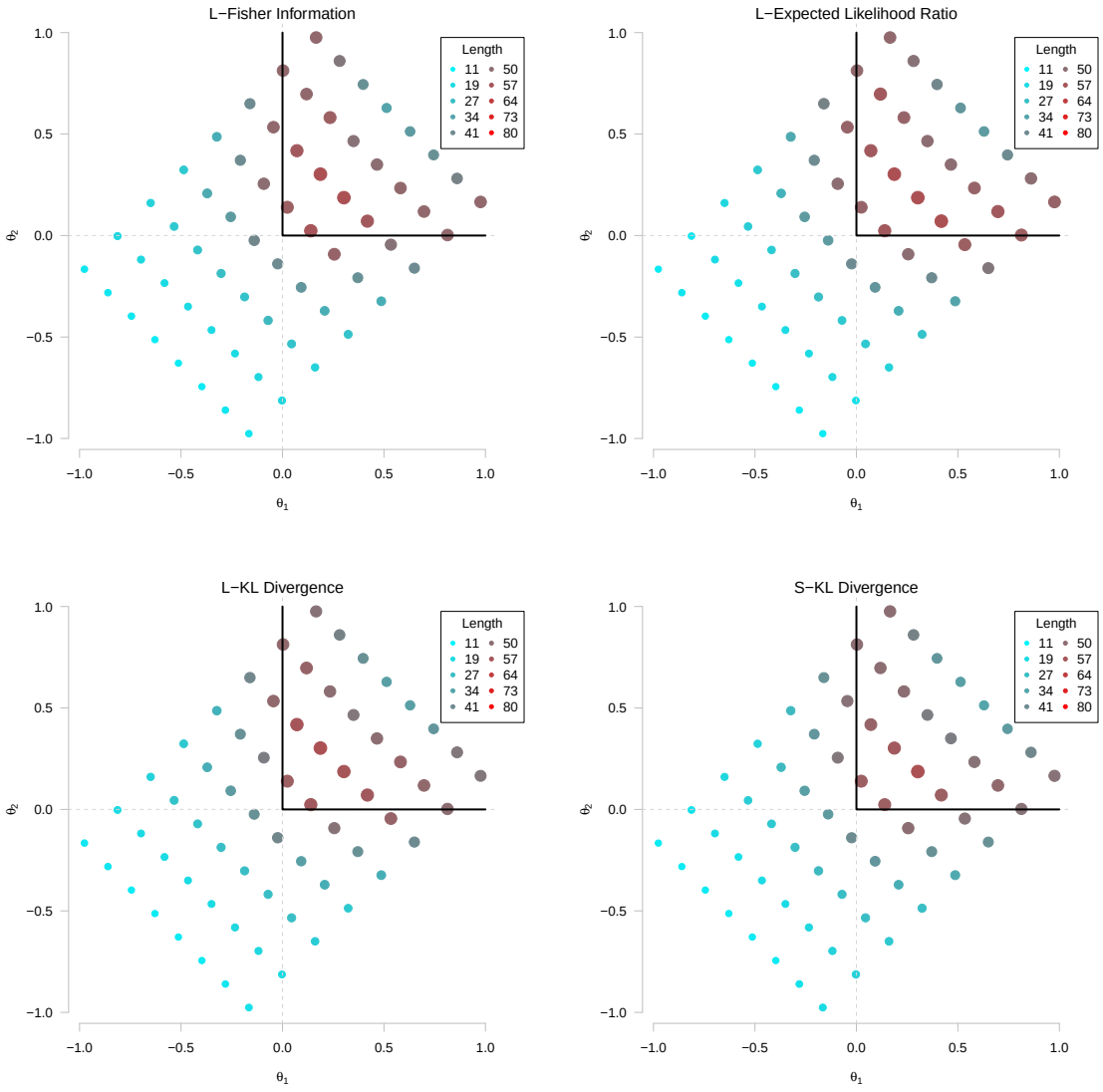

Figure 6.18: Scatterplots of the conditional average test length for various vectors of true ability when using the non-compensatory classification bound function and the BCR stopping rule with $\alpha = .10$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

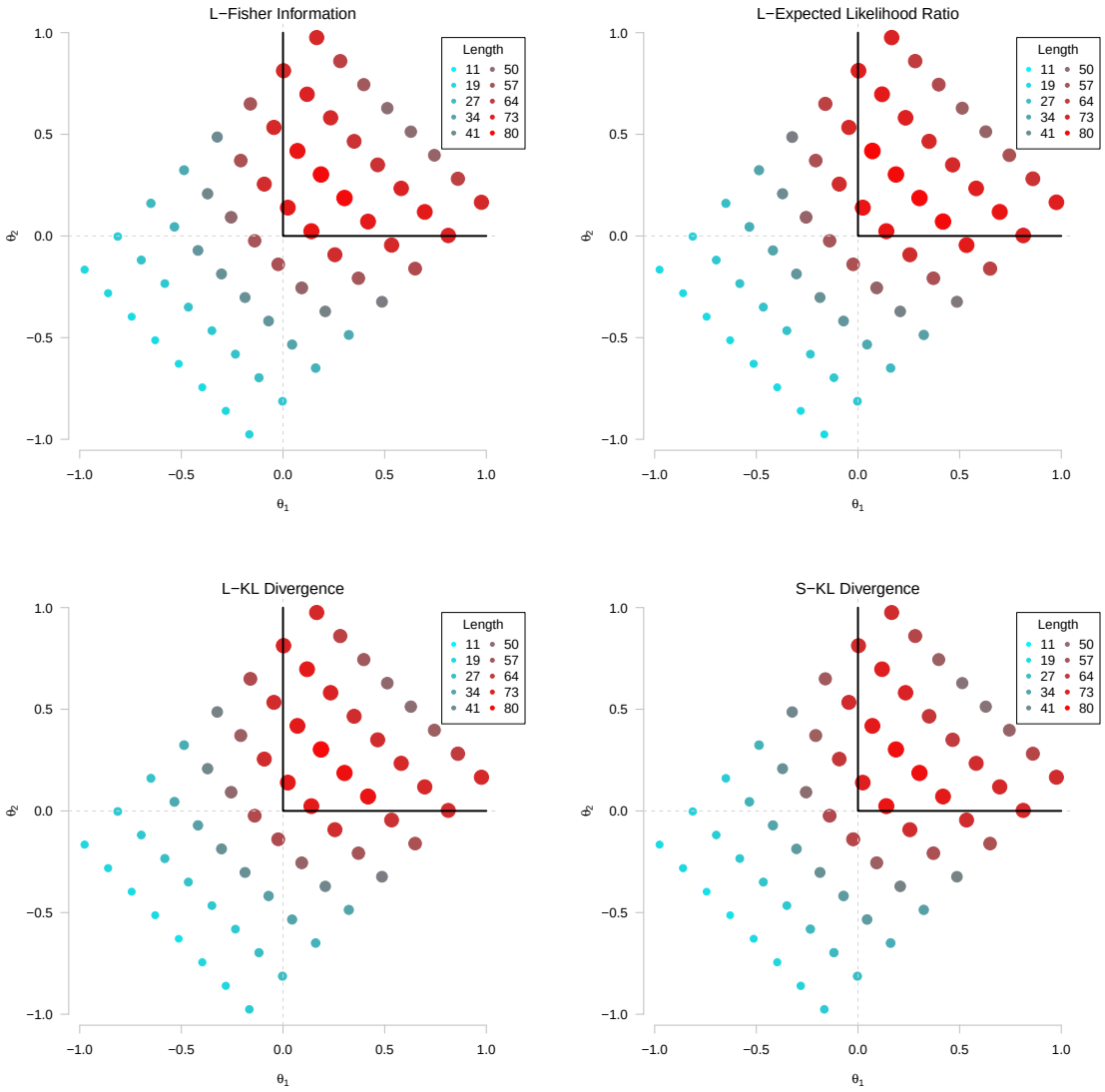

Figure 6.19: Scatterplots of the conditional average test length for various vectors of true ability when using the non-compensatory classification bound function and the BCR stopping rule with $\alpha = .05$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

The remaining plots conditional on a particular ability vectors, including the loss plots with $P = 500$, are all provided in Appendix D. Note that accuracy generally outweighs test length for the difficult-to-classify points near the classification bound function. Therefore, the loss plots generally appear to be similar to the classification accuracy plots, in that one finds little differences between item selection algorithms (with the exception of M-SCSPRT and L-ELR when using a non-compensatory classification bound) and minor differences among various stopping rules.

Chapter 7

Discussion and Conclusion

7.1 Summary and Discussion 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 categories when using a multidimensional IRT model. Conceptualizing multidimensional mastery testing requires constructing a mastery area that is separated from a non-mastery area by a classification bound function. Whereas unidimensional ability estimates are always a certain distance from the classification bound, multidimensional ability estimates can be located different distances from an infinite set of points on the classification bound function. Therefore, generalizing point-based comparisons, such as those required for the SPRT, necessitate choosing the appropriate "closest" point(s) on this function. One could define closest by those points that are closest in distance or those points that are closest in likelihood. Perhaps unsurprisingly, methods determining closest by the distance between an ability estimate and the classification bound function resulted in adequate performance only when the classification bound function aligned

with the contours of the likelihood function. In those instances, P-SPRT (closest in distance) resulted in similar loss to C-SPRT (closest in likelihood), primarily because the orthogonal projection from $\hat{\boldsymbol{\theta}}_i$ to $\boldsymbol{\theta}_0$ terminated in a $\hat{\boldsymbol{\theta}}_0$ that approximately maximized the constrained likelihood. If using a non-compensatory classification bound function, then the P-SPRT stopping rule was always less accurate than stopping rules with a similar average test length. In all cases, BCR yielded short and relatively accurate tests, and S-KL balanced efficiency with protecting against items chosen in uninformative directions.

As previously discussed, methods of improving stopping rules and item selection algorithms could include weighting individual test statistics by posterior densities and then averaging across some slice or region of the posterior distribution. This idea was applied to one of the tested stopping rules, BCR, and one of the tested item selection algorithms, S-KL. BCR was motivated by Berger's (2012) contention that posterior probabilities do not depend on the reason for stopping sequential tests. Because of the direct quantification of posterior probabilities, BCR resulted in the largest dependence upon testing attributes, such as the nominal α rate, compared to alternate stopping rules. Yet even if one were to eliminate those simulees who took tests of maximal length, the average classification accuracy of BCR did not approximate the nominal $1 - \alpha$ specification. Therefore, the stochastic dependence of a selected item on the responses to previously selected items must somewhat influence actual error rates for all of the stopping rules. A future study could compare the classification accuracy and test length when directly attempting to either quantify the error rates (as done in this study) or measure loss (as attempted by Glas & Vos, 2010). Glas and Vos (2010) explained that if minimizing the expected loss, then additional cost parameters could be added to capture model constraints, such as content balancing, exposure control, etc. However, Glas and Vos (2010) derived their stopping rules for a simplistic, Rasch-based

model, and any researchers attempting to further study their method would need to generalize their equations to more complicated multidimensional models.

With respect to item selection, S-KL yielded tests with approximately the best loss for either a compensatory or non-compensatory classification problem when using the most appropriate item bank for that problem. Classifying examinees with respect to a compensatory classification bound function was best when using the within-item multidimensional bank, classifying examinees with respect to a non-compensatory classification bound function was best when using the between-item multidimensional bank, and S-KL resulted in suitable loss in each of these cases. S-KL performed well even though the fixed-point item selection algorithm on which it was based, L-KL, resulted in relatively inaccurate and inefficient classifications. Yet both KL algorithms circumvented problems associated with the multidimensional KL indices that were discussed by Wang and Chang (2011) and Wang, Chang, and Boughton (2011). Wang, Chang, and Boughton (2011) explained that their multidimensional KL divergence index will select items that manifest large MDISC_j and where the difficulty of an item is similar to a linear combination of the current ability estimates (see Wang et al., p. 34). As shown in Wang and Chang (2011), these facts imply that multidimensional KL divergence will often select items that poorly differentiate the current ability estimate from true ability. However, the current study estimated true ability only with respect to the category in which that ability was located. Rather than integrating KL divergence across a region of multidimensional space, this study simply compared those points that were required for adequate classification. And by contrasting points along a line normal to the classification bound function, the KL indices discussed herein yield large values only if an item adequately differentiates masters from non-masters.

A common criticism of KL indices as applied to classification testing is that they

assume that every examinee is a priori in the mastery region. Therefore, maximizing KL divergence should result in less efficient classification tests if most examinees are in the non-mastery region. This characteristic of KL divergence was shown in the second simulation of Chapter 3, whereby the KL divergence-based item selection algorithm yielded the longest unidimensional classification tests when the cut-point was greater than the prior distribution mean. Unfortunately, the best unidimensional mastery testing item selection algorithm, ELR, which optimizes the expected log-likelihood ratio taken conditional on the current ability estimate, did not yield accurate multidimensional classifications in many cases. One could improve KL-based measures to somewhat approximate the decision-making process of ELR by taking into consideration the current ability estimate: if $\log\left[\mathrm{LR}(\hat{\boldsymbol{\theta}}_u, \hat{\boldsymbol{\theta}}_l|\mathbf{y}_i)\right] > 0$, maximize $\mathrm{KL}_j(\hat{\boldsymbol{\theta}}_u||\hat{\boldsymbol{\theta}}_l)$, but if $\log\left[\mathrm{LR}(\hat{\boldsymbol{\theta}}_u, \hat{\boldsymbol{\theta}}_l|\mathbf{y}_i)\right] < 0$, maximize $\mathrm{KL}_j(\hat{\boldsymbol{\theta}}_l||\hat{\boldsymbol{\theta}}_u)$. Future studies should consider this modification of the current KL-based item selection algorithms in both unidimensional and multidimensional mastery testing.

Several of the stopping rules and item selection algorithms could yield improved classification accuracy and test length if adapted to aggregate information across a distribution of simulees. For instance, stochastic curtailment resulted in poorly performing adaptive testing algorithms for both the compensatory and non-compensatory classification bound functions. With respect to the compensatory classification bound function, M-SCSPRT barely improved over C-SPRT in terms of average test length and classification accuracy when using the within-item multidimensional bank. But the loss function of M-SCSPRT essentially aligned with that of C-SPRT when using a compensatory classification bound function, so that test practitioners would gain very little over the simple C-SPRT procedure. With respect to the non-compensatory classification bound function, M-SCSPRT resulted in shorter test lengths but much worse classification accuracy than the C-SPRT procedure.

An alternative to classic stochastic curtailment is the Bayesian-like predictive power formation, described by Jennison and Turnbull (2000), applied by Finkelman (2010) to mastery tests, and presented in Equation (4.17) for multidimensional classification problems. Many authors have criticized predictive power because "it does not have a clear frequentist interpretation and at the same time is inconsistent with the principles of Bayesian theory" (Dmitrienko & Wang, 2006, p. 2179). However, as described by Finkelman (2010), "predictive power implicitly takes the uncertainty about θ into account via the posterior distribution" (p. 36), which should reduce the dependence of the conditional probabilities on a particular set of items chosen assuming a particular $\boldsymbol{\theta}_i$. One might also improve C-SPRT and M-GLR by averaging across either the posterior distribution (as described by Dickey, 1971 and Kharin, 2011, and presented in modified form as W-GLR in Equation 4.12) or along the classification bound function. Although the C-SPRT algorithm compares points of similar likelihood on both sides of the classification bound function, the ratio of those values might be abnormally small relative to the ratio of likelihoods a short distance along the classification bound function away. Moreover, by directly quantifying the average of some attribute within a region, aggregative methods should also result in adequate error rates for more complicated classification problems, such as those with irregularly shaped classification bound functions or more than two categories.

Practitioners should also consider the computation time of various item selection algorithms and stopping rules. Even though S-KL and BCR resulted in tests with the best loss in many circumstances, they required 6–18 times as much computing time using `R` (R Development Core Team, 2013) on a 2GHz Intel Core i7 processor. Many of the promising item selection algorithms or stopping rules could not be tested in this study due to the excessively long computing times required. Programmers could design more efficient algorithms by writing all code directly in a compiled language, such as C,

but they must know optimal methods of estimating multidimensional integrals to derive any computing benefit.

7.2 Conclusion

Classification tests can be used to assess everything from job qualifications (e.g., ACT, 2007) to teacher certification (e.g., Pearson, 2011). Test items inherently assess more than one dimension (e.g., Ackerman, Gierl, & Walker, 2003), but practitioners have few methods available for multidimensional classification. One could, of course, treat these multidimensional traits as unidimensional and proceed with well-understood computerized mastery testing algorithms. One could also assume that the test measures a constellation of discrete states and adopt a popular Cognitive Diagnosis CAT algorithm (e.g., Cheng, 2009; Wang, Chang, & Douglas, 2011). However, most large-scale tests are based on IRT, and neither of the alternative methods allow for complex areas of classification. Future researchers would be better served by applying the various array of sequential tests developed in the statistical and biological literature (e.g., Bartroff & Lai, 2010; Dallow & Fina, 2010; Dmitrienko & Wang, 2006; Kharin, 2011; Todd, 2007) to psychometric questions.

This thesis is an attempt to explore the questions associated with applying sequential analysis to multidimensional psychometric problems. Future studies should derive and compare alternate generalizations of SPRT and GLR to multidimensional classification tasks with a variety of item banks and classification bound functions. Many of the methods discussed herein have great promise to control the misclassification rate of multidimensional classification tests. Psychometricians could better direct these methods to yield adequate classifications by constructing appropriate regions of integration and/or choosing more appropriate critical values. Due to a maximum test length in

all CATs, any generalization of sequential stopping rules to multiple dimensions must include associated curtailment methods. One such method is strict stochastic curtailment (referred to as "conditional power" by Dmitrienko & Wang, 2006), but researchers have also discussed "predictive power" (combining Bayesian and frequentist methods) and "predictive probability" (using a strictly Bayesian framework) approaches. Because of the high costs of computation in multiple dimensions, researchers must consider the computation time when providing recommendations.

To best apply classification algorithms to multiple dimensions, psychometricians must understand how those algorithms perform for mastery tests comprised of only one trait. Much of this thesis discussed open questions in unidimensional classification testing, including methods of finding the optimal difficulty parameter for efficient classification or improving previously existing item selection algorithms by considering the estimated location of the examinee with respect to the classification bound function. Other existing research topics include applying currently existing algorithms to classification problems with more than two categories. Several studies have extended stochastic curtailment to multiple categories (Wouda & Eggen, 2009) or polytomous IRT models (Gnambs & Batnic, 2011). However, the authors of these studies used stopping rules that were arbitrary extensions of Finkelman (2008a) to more than two categories and alternate models. Better understanding the performance of curtailment-based methods when using unidimensional models would improve generalizations of those algorithms to multiple dimensional classification problems.

In practice, items should be chosen from banks that are based off of realistic exams. Many authors (e.g., Bartroff et al., 2008; Finkelman, 2008a; Gnambs & Batnic; Wouda & Eggen, 2009) have used items from either traditional achievement or personality measures. The current study simulated item banks to retain certain distributional properties. These artificial banks were constructed to reduce dependence of item selection

algorithm and stopping rule performance on a specific set of limited items. However, all methods, no matter how abstract and preliminary, must eventually be applied to serviceable exams.

Although this project varied item selection algorithms and stopping rules, other properties of the testing algorithm, such as ability estimation, are important in the performance of adaptive tests. The current study adopted a modified MLE algorithm due to time constraints, but Bayesian methods should yield stabler estimates of ability earlier in an adaptive test. Other authors have extended bias-reduction methods, such as Warm's (1989) WLE method, to estimate a multidimensional ability vector (e.g., Tseng & Hsu, 2001; Wang, 2013). Despite focusing on classification, several mastery testing stopping rules depend on adequate estimation of examinee ability. Moreover, simulations in Chapter 3 showed that including knowledge of the current ability estimate results in a more efficient item selection algorithm than only considering the classification bound. Therefore, future research must also examine how various ability estimators interact with stopping rules and item selection algorithms to yield efficient and accurate classifications.

The discussions and simulations presented herein demonstrate the applicability of sequential algorithms to classify examinees in regions of multidimensional space. With the exception of determining the closest classification bound via projection, all of the stopping rules resulted in fairly efficient and accurate multidimensional mastery tests. Limitations of this study, of course, include the small number of conditions and the assumed latent space of only $K = 2$ dimensions. These methods should easily extend to $K \geq 3$ dimensions, but the computing time required as K increases would quickly become exceedingly long. Future research should adopt and extend the proposed algorithms to supply adaptive mastery tests that deliver the most accurate classifications

using the least number of items. Only after calibrating the properties of classification algorithms to exact specifications can practitioners use these algorithms to make informative and productive decisions.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, *97*, 129–133.

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, *13*, 113–127.

Ackerman, T. A. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement*, *18*, 257–275.

Ackerman, T., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational psychological tests. *Educational Measurement: Issues and Practices*, *22*, 37–51.

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

Babcock, B. (2011). Estimating a noncompensatory IRT model using Metropolis within Gibbs sampling. *Applied Psychological Measurement*, *35*, 317–329.

Babcock, B. (2009). Termination criteria in computerized adaptive test: Variable length tests are not biased. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference*

on computerized adaptive testing. Retrieved June 29, 2011 from: www.psych.umn.edu/psylabs/CATCentral

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

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

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

Berger, J. (2012). *Lecture 2: Bayesian hypothesis testing*. Presented at the CBMS Conference on Model Uncertainty and Multiplicity, Santa Cruz, CA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, *27*, 395–414.

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

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

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

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

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

Cheng, P. E., & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement*, *24*, 257–265.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing; CD-CAT. *Psychometrika*, *74*, 619–632.

Dallow, N., & Fina, P. (2010). The perils with the misuse of predictive power. *Pharmaceutical Statistics*, *2011*, 311–317.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *42*, 204–223.

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

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

Eggen, T. J. H. M. (2010). Three category adaptive classification testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing*. New York, NY: Springer.

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

Eisenberg, B., Gosh, B. K., & Simons, G. (1976). Propoerties of generalized sequential probability ratio tests. *Annals of Statistics*, *4*, 237–251.

Eisenberg, B., & Simons, G. (1978). On weak admissibility of tests. *Annals of Statistics*, *6*, 319–332.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika*, *49*, 175–186.

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

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

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

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

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

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

Frey, A., & Seitz, N.-N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*, *35*, 89–94.

Gierl, M. J., & Zhou, J. (2008). Computer adaptive-attribute testing: A new approach to cognitive diagnostic assessment. *Zeitschrift für Psychologie*, *216*, 29–39.

Glas, C. A. W., & Vos, H. J. (2010). Adaptive mastery testing using a multidimensional IRT model. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing*. New York, NY: Springer.

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

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

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

Interpreting scores (2013, winter). *Radiography: Certification handbook and application materials*. Retrieved from: https://www.arrt.org/pdfs/Disciplines/Handbooks/RAD-Handbook-new.pdf

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

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.

Kharin, A. (2011). Robustness analysis for Bayesian sequential testing of composite hypotheses under simultaneious distortion of priors and likelihoods. *Austrian Journal of Statistics*, *40*, 65–73.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237–254). New York: Academic Press.

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

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

Jha, S. K., Clarke, E. M., Langmead, C. J., Legay, A., Platzer, A., & Zuliani, P. (2009). A Bayesian approach to model checking biological systems. In P. Degano & R. Gorrieri (Eds.), *Proceedings of the 7th international Computational Methods in Systems Biology Conference* (pp. 218–234). Berlin: Springer Berlin Heidelberg.

Lachin, J. M. (1981). Sequential clinical trials for normal variates using interval composite hypotheses. *Biometrics*, *37*, 87–101.

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

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

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

Lavine, M. & Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *The American Statistician*, *53*, 119–122.

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

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

Lin, C.-J., & Spray, J. A. (2000). *Effects of item-selection criteria on classification testing with the sequential probability ratio test* (Research Report No. 2000-8). Iowa City, IA: ACT.

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

McGlohen, M. & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, *40*, 808–821.

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

Mulder, J., & van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback-Leibler information item selection. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing.* New York, NY: Springer.

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

Nydick, S. W. (2012). *Accuracy and efficiency in classifying examinees using computerized adaptive tests* (Unpublished master's thesis). University of Minnesota, Minneapolis, MN.

Nydick, S. W. (2013). catIrt: An R package for simulating computerized adaptive tests. R package version 0.4-1.

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

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

R Core Team. (2013). R: A language and environment for statistical computing (Version 3.0.1) Vienna, Austria: R Foundation for Statistical Computing. URL: http://www.R-project.org/.

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. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401–412.

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

Roussos, L., DiBello, L. V., Stout, W., Hartz, S., Henson, R. A., & Templin, J. H. (2007). *The fusion model skills diagnosis system*. In J. P. Leighton, & Gierl, M. J. (Eds.), *Cognitively diagnostic assessment for education: Theory and practice*. (pp. 275–318). Thousand Oaks, CA: SAGE.

Rudner, L. M. (2009). An examination of decision-theory adaptive testing procedures. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*. Retrieved July 1, 2011 from: www.psych.umn.edu/psylabs/CATCentral

Rupp, A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*, 219–262.

Rupp, A., Templin, J., & Hensen, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: The Guilford Press.

Seitz, N.-N., & Frey, A. (2013). The sequential probability ratio test for multidimensional adaptive testing with between-item multidimensionality. *Psychological Test and Assessment Modeling*, *55*, 105–123.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354.

Segall, D. O. (2000). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53-57). Dordrecht, NL: Kluwer.

Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika*, *66*, 79–97.

Seitz, N.-N., & Frey, A. (2013). The sequential probability ratio test for multidimensional adaptive testing with between-item multidimensionality. *Psychological Test and Assessment Modeling*, *55*, 105–123.

Savalei, V. (2006). Logistic approximation to the normal: The KL rationale. *Psychometrika*, *71*, 763–767.

Smarter Balanced Assessment Consortium. (2013). *Smarter Balanced Milestones*. Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2013/04/Smarter-Balanced-Milestones.pdf

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

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

Spray, J. A., Abdel-fatah, A. A., Huang, C.-Y., & Lau, C. A. (1997). *Unidimensional approximations for a computerized test when the item pool and latent space are multidimensional*. (Research Report No.97-5). Iowa City, IA: ACT.

Stewart, J. (2007). *Essential calculus: Early transcendentals*. Belmont, CA: Thompson Brooks/Cole.

Thompson, N. A. (2009). Using the generalized likelihood ratio as a termination criterion. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*. Retrieved June 29, 2011 from: www.psych.umn.edu/psylabs/CATCentral

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

Todd, S. (2007). A 25-year review of sequential methodology in clinical studies. *Statistics in Medicine*, *26*, 237–252.

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, *24*, 398–412.

van der Linden, W. J. (2012). On compensation in multidimensional response modeling. *Psychometrika*, *77*, 21–30.

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

Vos, H. J. (1999). Applications of Bayesian decision theory to sequential mastery testing. *Journal of Educational and Behavioral Statistics*, *24*, 271–292.

Vos, H. J. (2002). Applying the minimax principle to sequential mastery testing. In A. Ferligoj and A. Mrvar (Eds.), *Developments in Social Science Methodology*. Metodoloski zvezki, 18, Ljubljana: FDV.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*, 307–333.

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

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

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

Wang, C. On latent trait estimation in mutldimensional compensatory item response models. Manuscript submitted for publication.

Wang, C., & Chang, H.-H. (2011). Item selection in multidimensional computerized adaptive testing—gaining information from different angles. *Psychometrika*, *76*, 363–384.

Wang, C., Chang, H.-H., & Douglas, Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavioral Research*, *44*, 95–109.

Wang, C., Chang, H.-H., & Boughton, K. A. (2011). Kullback-Leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika*, *76*, 13–39.

Wang, C., Chang, H.-H., & Huebner, A. (2011). Restrictive item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, *48*, 255–273.

Wang, W.-C., & Huang, S.-Y. (2011). Computerized classification testing under the one-parameter logistic response model with ability-based guessing. *Educational and Psychological Measurement*, *71*, 925–941.

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *28*, 450–480.

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

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

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

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

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

Wiberg, M. (2003). An optimal design approach to criterion-referenced computerized testing. *Journal of Educational and Behavioral Statistics*, *28*, 97–110.

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

Appendix A

Derivations

A.1 Maximum of the Log-Likelihood Ratio for a Correct Response

To determine the effect of item parameters on the classification bound (θ_0) yielding the best evidence for classification, assume an examinee correctly responded to one 3PL item with arbitrary a (discrimination), b (difficulty), c (guessing), and δ (half-width of the indifference region) parameters. Then we can define a function

$$f_1(\theta_0) = \log\left[\frac{p(\theta_0+\delta)}{p(\theta_0-\delta)}\right] = \log[p(\theta_0+\delta)] - \log[p(\theta_0-\delta)], \tag{A.1}$$

where $p(x) = c + (1-c)\frac{\exp[a(x-b)]}{1+\exp[a(x-b)]}$ is the item response function (IRF) for the three-parameter logistic model (3PL), and a, c, θ, θ_0, and δ are fixed/constant parameters. Determining θ_0 such that $f_1(\theta_0)$ is at a maximum results in a classification bound providing the best evidence that an examinee is above it. Therefore, we should take the derivative of Equation (A.1) and set it equal to 0.

$$\begin{aligned}\frac{df_1(\theta_0)}{d\theta_0} &= \left[\frac{1}{p(\theta_0+\delta)}\right]\left[\frac{dp(\theta_0+\delta)}{d\theta_0}\right] - \left[\frac{1}{p(\theta_0-\delta)}\right]\left[\frac{dp(\theta_0-\delta)}{d\theta_0}\right] \\ &= \left[\frac{1}{c+(1-c)\frac{\exp[a(\theta_0+\delta-b)]}{1+\exp[a(\theta_0+\delta-b)]}}\right]\left[\frac{dp(\theta_0+\delta)}{d\theta_0}\right] - \left[\frac{1}{c+(1-c)\frac{\exp[a(\theta_0-\delta-b)]}{1+\exp[a(\theta_0-\delta-b)]}}\right]\left[\frac{dp(\theta_0-\delta)}{d\theta_0}\right]\end{aligned} \tag{A.2}$$

But $\frac{dp(x)}{dx} = \frac{(1-c)a\exp[a(x-b)]}{(1+\exp[a(x-b)])^2}$, and after simplifying, we have

$$\frac{df_1(\theta_0)}{d\theta_0} = \left[\frac{(1-c)a}{c+\exp[a(\theta_0+\delta-b)]}\right]\left[\frac{\exp[a(\theta_0+\delta-b)]}{1+\exp[a(\theta_0+\delta-b)]}\right] - \left[\frac{(1-c)a}{c+\exp[a(\theta_0-\delta-b)]}\right]\left[\frac{\exp[a(\theta_0-\delta-b)]}{1+\exp[a(\theta_0-\delta-b)]}\right]. \tag{A.3}$$

To find the maximum of the log-likelihood ratio set Equation (A.3) equal to 0 and solve for θ_0.

$$\begin{aligned}\left[\frac{\exp[a(\theta_0+\delta-b)]}{\exp[a(\theta_0-\delta-b)]}\right]\left[\frac{1+\exp[a(\theta_0-\delta-b)]}{1+\exp[a(\theta_0+\delta-b)]}\right] &= \frac{c+\exp[a(\theta_0+\delta-b)]}{c+\exp[a(\theta_0-\delta-b)]} \\ \exp(2a\delta)\left[\frac{1+\exp[a(\theta_0-\delta-b)]}{1+\exp[a(\theta_0+\delta-b)]}\right] &= \frac{c+\exp[a(\theta_0+\delta-b)]}{c+\exp[a(\theta_0-\delta-b)]}.\end{aligned} \tag{A.4}$$

Let $\gamma = \exp(2a\delta)$ in Equation (A.4) for simplicity. Then

$$\gamma(1+\exp[a(\theta_0-\delta-b)])(c+\exp[a(\theta_0-\delta-b)]) = (1+\exp[a(\theta_0+\delta-b)])(c+\exp[a(\theta_0+\delta-b)]). \tag{A.5}$$

The left side of Equation (A.5) becomes

$$\begin{aligned}
&\gamma[(1+\exp[a(\theta_0-\delta-b)])(c+\exp[a(\theta_0-\delta-b)]) \\
&\quad = \gamma\left[c + c\exp[a(\theta_0-\delta-b)] + \exp[a(\theta_0-\delta-b)] + \exp[a(2\theta_0-2\delta-2b)]\right] \\
&\quad = c\gamma + c\gamma\exp[a\theta_0]\exp[-a\delta]\exp[-ab] + \gamma\exp[a\theta_0]\exp[-a\delta]\exp[-ab] + \gamma\exp[a(2\theta_0-2\delta-2b)] \\
&\quad = c\exp[2a\delta] + c\exp[a\theta_0]\exp[a\delta]\exp[-ab] + \exp[a\theta_0]\exp[a\delta]\exp[-ab] + \exp[2a\theta_0]\exp[-2ab].
\end{aligned} \tag{A.6}$$

And the right side of Equation (A.5) becomes

$$\begin{aligned}
&(1+\exp[a(\theta_0-\delta-b)])(c+\exp[a(\theta_0-\delta-b)]) \\
&\quad = c + c\exp[a(\theta_0+\delta-b)] + \exp[a(\theta_0+\delta-b)] + \exp[a(2\theta_0+2\delta-2b)] \\
&\quad = c + c\exp[a\theta_0]\exp[a\delta]\exp[-ab] + \exp[a\theta_0]\exp[a\delta]\exp[-ab] + \exp[2a\theta_0]\exp[2a\delta]\exp[-2ab].
\end{aligned} \tag{A.7}$$

The middle two terms in Equations (A.6) and (A.7) are identical, so they cancel, and Equation (A.5) simplifies to

$$\begin{aligned}
c\exp[2a\delta] + \exp[2a\theta_0]\exp[-2ab] &= c + \exp[2a\theta_0]\exp[2a\delta]\exp[-2ab] \\
\exp[2a\theta_0]\exp[-2ab](1-\exp[2a\delta]) &= c(1-\exp[2a\delta]) \\
\exp[2a\theta_0]\exp[-2ab] &= c \\
\exp[2a\theta_0] &= \frac{c}{\exp[-2ab]}.
\end{aligned} \tag{A.8}$$

Finally, taking logs of both sides and dividing, we find that the maximum of $f(\theta_0)$ with respect to θ_0 is

$$\log(\exp[2a\theta_0]) = \log\left(\frac{c}{\exp[-2ab]}\right)$$
$$2a\theta_0 = \log(c) + 2ab$$
$$\hat{\theta}_0 = \frac{\log(c)}{2a} + b. \tag{A.9}$$

Therefore, given one item with parameters a, b, and c, the classification bound that maximizes the log-likelihood ratio (assuming a correct response) is $\theta_0 = \frac{\log(c)}{2a} + b$. As $c \to 0$, then $\log(c) \to -\infty$, so that the evidence for classification is a monotone function of the classification bound, but as $c \to 1$, then the classification bound that maximizes the log-likelihood ratio approaches $\theta_0 = b$.

A.2 Maximum of the Expected Log-Likelihood Ratio with respect to θ_0

If we do not assume that an examinee responded correctly to an item, the appropriate objective function is the *expected* log-likelihood ratio,

$$f_2(\theta_0) = p(\theta)\Big(\log[p(\theta_0+\delta)] - \log[p(\theta_0-\delta)]\Big) + q(\theta)\Big(\log[q(\theta_0+\delta)] - \log[q(\theta_0-\delta)]\Big), \quad \text{(A.10)}$$

where $p(x) = c + (1-c)\frac{\exp[a(x-b)]}{1+\exp[a(x-b)]}$ is the item response function (IRF) for the three-parameter logistic model (3PL), $q(x) = 1-p(x)$, and a, c, θ, θ_0, and δ are fixed/constant parameters. To find θ_0 such that $f_2(\theta_0)$ is at a maximum, take the derivative of Equation (A.10) and set it equal to 0. The derivative of Equation (A.10) can be written as

$$\frac{df_2(\theta_0)}{d\theta_0} = ap(\theta)[p^c(\theta_0+\delta) - p^c(\theta_0-\delta)] - a[p^1(\theta_0+\delta) - p^1(\theta_0-\delta)] \quad \text{(A.11)}$$

where $p^1(x) = \frac{\exp[a(x-b)]}{1+\exp[a(x-b)]}$ and $p^c(x) = \frac{\exp[a(x-b)]}{c+\exp[a(x-b)]}$. After finding a common denominator for the first half of the right side and the left half of the right side of Equation (A.11), we have

$$\begin{aligned}\frac{df_2(\theta_0)}{d\theta_0} = {} & ac\left[\frac{\exp[a(\theta_0+\delta-b)] - \exp[a(\theta_0-\delta-b)]}{c^2 + c\exp[a(\theta_0+\delta-b)] + c\exp[a(\theta_0-\delta-b)] + \exp[2a(\theta_0-b)]}\right] \\ & - ap(\theta)\left[\frac{\exp[a(\theta_0+\delta-b)] - \exp[a(\theta_0-\delta-b)]}{1 + \exp[a(\theta_0+\delta-b)] + \exp[a(\theta_0-\delta-b)] + \exp[2a(\theta_0-b)]}\right]. \quad \text{(A.12)}\end{aligned}$$

To find the maximum of the expected log-likelihood ratio, set Equation (A.12) equal to 0 and solve for θ_0. First note that we can divide out constants and simplify the

derivative to

$$\frac{1}{1+\exp[a(\theta_0+\delta-b)]+\exp[a(\theta_0-\delta-b)]+\exp[2a(\theta_0-b)]} = \frac{cp(\theta_i)}{c^2+c\exp[a(\theta_0+\delta-b)]+c\exp[a(\theta_0-\delta-b)]+\exp[2a(\theta_0-b)]}. \quad \text{(A.13)}$$

Next, cross-multiply fractional denominators, which results in

$$cp(\theta_i)+cp(\theta_i)\exp[a(\theta_i+\delta-b)]+cp(\theta_i)\exp[a(\theta_i-\delta-b)]+cp(\theta_i)\exp[2a(\theta_0-b)] = c^2+c\exp[a(\theta_0+\delta-b)]+c\exp[a(\theta_0-\delta-b)]+\exp[2a(\theta_0-b)]. \quad \text{(A.14)}$$

After applying multiplicative properties of the exponential function, gathering terms that contain θ_0, and dividing all of the terms by $cp(\theta_i)-1$, we are left with

$$\exp[2a(\theta_0-b)]+\exp[a(\theta_0-b)]\left(\exp[a\delta]+\exp[-a\delta]\right)\left(\frac{cp(\theta_i)-c}{cp(\theta_i-1)}\right) = \frac{c^2-cp(\theta_i)}{cp(\theta_i)-1}. \quad \text{(A.15)}$$

Note that Equation (A.15) can be written as $x^2+x\gamma=\psi$, where $x=\exp[a(\theta_0-b)]$, $\psi=\frac{c^2-cp(\theta_i)}{cp(\theta_i)-1}$, and $\gamma=\left(\exp[a\delta]+\exp[-a\delta]\right)\left(\frac{cp(\theta_i)-c}{cp(\theta_i-1)}\right)$. Applying the property that if $x^2+x\gamma=\psi$, then $x=\sqrt{\psi+\frac{\gamma^2}{4}}-\frac{\gamma}{2}$ to Equation (A.15), we have

$$\begin{aligned}\exp[a(\theta_0-b)] &= \left(\frac{c^2-cp(\theta_i)}{cp(\theta_i)-1}+\frac{1}{4}\left[\left(\exp[a\delta]+\exp[-a\delta]\right)\left(\frac{cp(\theta_i)-c}{cp(\theta_i)-1}\right)\right]^2\right)^{1/2} \\ &\qquad -\frac{1}{2}\left(\exp[a\delta]+\exp[-a\delta]\right)\left(\frac{cp(\theta_i)-c}{cp(\theta_i)-1}\right) \\ &= \left(\frac{c^2-cp(\theta_i)}{cp(\theta_i)-1}+\cosh^2[a\delta]\left(\frac{cp(\theta_i)-c}{cp(\theta_i)-1}\right)^2\right)^{1/2}-\cosh[a\delta]\left(\frac{cp(\theta_i)-c}{cp(\theta_i)-1}\right), \quad \text{(A.16)}\end{aligned}$$

which uses the identity $\cosh(x) = \frac{\exp(x)+\exp(-x)}{2}$. Equation (A.16) can be simplified further by expanding $p(\theta_i)$. There are ultimately *three* terms in Equation (A.16) that include $p(\theta_i)$:

1. $c^2 - cp(\theta_i)$
2. $cp(\theta_i) - c$
3. $cp(\theta_i) - 1$

We can simplify each of those terms in turn. The first term simplifies to

$$\begin{aligned} c^2 - cp(\theta_i) &= c^2 - c\left[c + (1-c)\left(\frac{\exp[a(\theta_i - b)]}{1+\exp[a(\theta_i - b)]}\right)\right] \\ &= c(c-1)\left(\frac{\exp[a(\theta_i - b)]}{1+\exp[a(\theta_i - b)]}\right). \end{aligned} \tag{A.17}$$

The next term simplifies to

$$\begin{aligned} cp(\theta_i) - c &= c[p(\theta_i) - 1] \\ &= c\left[c + (1-c)\frac{\exp[a(\theta_i - b)]}{1+\exp[a(\theta_i - b)]} - 1\right] \\ &= c\left[\frac{c + c\exp[a(\theta_i - b)]}{1+\exp[a(\theta_i - b)]} + \frac{\exp[a(\theta - b)] - c\exp[a(\theta_i - b)]}{1+\exp[a(\theta_i - b)]} - \left(\frac{1+\exp[a(\theta_i - b)]}{1+\exp[a(\theta_i - b)]}\right)\right] \\ &= \frac{c(c-1)}{1+\exp[a(\theta_i - b)]}. \end{aligned} \tag{A.18}$$

And the final term simplifies to

$$
\begin{aligned}
cp(\theta_i) - 1 &= c\left[c + (1-c)\frac{\exp[a(\theta_i - b)]}{1+\exp[a(\theta_i - b)]}\right] - 1 \\
&= \frac{c^2 + c^2\exp[a(\theta_i - b)]}{1+\exp[a(\theta_i - b)]} + \frac{c\exp[a(\theta_i - b)] - c^2\exp[a(\theta_i - b)]}{1+\exp[a(\theta_i - b)]} - \left(\frac{1+\exp[a(\theta_i - b)]}{1+\exp[a(\theta_i - b)]}\right) \\
&= \frac{(c^2-1) + (c-1)\exp[a(\theta_i - b)]}{1+\exp[a(\theta_i - b)]}. \qquad \text{(A.19)}
\end{aligned}
$$

Replacing Equations (A.17), (A.18), and (A.19) in the fractions of Equation (A.16) containing those terms, we have

$$
\begin{aligned}
\frac{c^2 - cp(\theta_i)}{cp(\theta_i) - 1} &= \frac{c(c-1)\exp[a(\theta_i - b)]}{(c^2-1) + (c-1)\exp[a(\theta_i - b)]} \\
&= \frac{c(c-1)\exp[a(\theta_i - b)]}{(c-1)(c+1) + (c-1)\exp[a(\theta_i - b)]} \\
&= \frac{c\exp[a(\theta_i - b)]}{c+1+\exp[a(\theta_i - b)]} \qquad \text{(A.20)}
\end{aligned}
$$

and we have

$$
\begin{aligned}
\frac{cp(\theta_i) - c}{cp(\theta_i) - 1} &= \frac{c(c-1)}{(c^2-1) + (c-1)\exp[a(\theta_i - b)]} \\
&= \frac{c}{c+1+\exp[a(\theta_i - b)]}. \qquad \text{(A.21)}
\end{aligned}
$$

After inserting Equations (A.20) and (A.21), Equation (A.16) simplifies to

$$\begin{aligned}\exp[a(\theta_0 - b)] &= \left(\frac{c\exp[a(\theta_i - b)]}{c + 1 + \exp[a(\theta_i - b)]} + \cosh^2[a\delta]\left[\frac{c^2}{(c + 1 + \exp[a(\theta_i - b)])^2}\right]\right)^{1/2} \\ &\qquad - \cosh[a\delta]\left[\frac{c}{c + 1 + \exp[a(\theta_i - b)]}\right] \\ &= \frac{\left[c^2(\exp[a(\theta_i - b)] + \cosh^2[a\delta]) + c(\exp[a(\theta_i - b)] + \exp[2a(\theta_i - b)])\right]^{1/2} - c\cosh[a\delta]}{c + 1 + \exp[a(\theta_i - b)]}\end{aligned} \tag{A.22}$$

Letting $g = \left[c^2(\exp[a(\theta_i - b)] + \cosh^2[a\delta]) + c(\exp[a(\theta_i - b)] + \exp[2a(\theta_i - b)])\right]^{1/2}$ and $h = c\cosh[a\delta]$, multiply the numerator and denominator of Equation (A.22) by $g + h$ to yield

$$\exp[a(\theta_0 - b)] = \frac{c\exp[a(\theta_i - b)]}{\left[c\exp[a(\theta_i - b)]\{c + 1 + \exp[a(\theta_i - b)]\} + \big(c\cosh[a\delta]\big)^2\right]^{1/2} + \big(c\cosh[a\delta]\big)}. \tag{A.23}$$

Finally, taking logs of both sides of Equation (A.23) and simplifying results in

$$\begin{aligned}\hat{\theta}_0 &= \frac{\log(c)}{a} + \theta_i - \frac{\log\left(\left[c\exp[a(\theta_i - b)]\{c + 1 + \exp[a(\theta_i - b)]\} + \big(c\cosh[a\delta]\big)^2\right]^{1/2} + \big(c\cosh[a\delta]\big)\right)}{a} \\ &= \frac{\log(c)}{2a} + \theta_i - \frac{\log\left(\left[\exp[a(\theta_i - b)]\{c + 1 + \exp[a(\theta_i - b)]\} + \big(c^{1/2}\cosh[a\delta]\big)^2\right]^{1/2} + \big(c^{1/2}\cosh[a\delta]\big)\right)}{a}.\end{aligned} \tag{A.24}$$

A.3 Maximum of the Expected Log-Likelihood Ratio with respect to b

One might, instead, seek to determine the optimal item difficulty, b, that minimizes (if $\theta < \theta_0$) or maximizes (if $\theta > \theta_0$) the expected log-likelihood ratio. The objective function is the same, only now a function of b rather than θ_0, namely

$$f_2(b) = p(\theta)\Big(\log[p(\theta_0+\delta)] - \log[p(\theta_0-\delta)]\Big) + q(\theta)\Big(\log[q(\theta_0+\delta)] - \log[q(\theta_0-\delta)]\Big), \quad \text{(A.25)}$$

where $p(x) = c + (1-c)\frac{\exp[a(x-b)]}{1+\exp[a(x-b)]}$ is the item response function (IRF) for the three-parameter logistic model (3PL), $q(x) = 1-p(x)$, and a, c, θ, θ_0, and δ are fixed/constant parameters. To find b such that $f_2(b)$ is at a minimum/maximum, take the derivative of Equation (A.25) and set it equal to 0. Using the product rule, this derivative can be broken down into two parts. First, note that holding $p(\theta)$ and $q(\theta)$ constant, the derivative is identical to that from the previous derivation with an extra negative out front due to the chain rule. Therefore, the derivative of Equation (A.25) (holding $p(\theta)$ and $q(\theta)$ constant) can be written as

$$\frac{df_2(b)}{db}^I = -ap(\theta)[p^c(\theta_0+\delta) - p^c(\theta_0-\delta)] - a[p^1(\theta_0+\delta) - p^1(\theta_0-\delta)], \quad \text{(A.26)}$$

where $p^1(x) = \frac{\exp[a(x-b)]}{1+\exp[a(x-b)]}$ and $p^c(x) = \frac{\exp[a(x-b)]}{c+\exp[a(x-b)]}$. Next, note that $\frac{dp(\theta)}{db} = -(1-c)ap^1(\theta)q^1(\theta)$ and $\frac{dq(\theta)}{db} = -\frac{dp(\theta)}{db} = (1-c)ap^1(\theta)q^1(\theta)$, so that the derivative of Equation (A.25) (holding $\{\log[p(\theta_0+\delta)] - \log[p(\theta_0-\delta)]\}$ and $\{\log[q(\theta_0+\delta)] - \log[q(\theta_0-\delta)]\}$ constant) can be written as

$$\frac{df_2(b)}{db}^{II} = -(1-c)ap^1(\theta)q^1(\theta)\log\left[\frac{p(\theta_0+\delta)q(\theta_0-\delta)}{q(\theta_0+\delta)p(\theta-\delta)}\right]. \tag{A.27}$$

The full derivative of Equation (A.25) with respect to b is simply the sum of Equations (A.26) and (A.27), or

$$\begin{aligned}\frac{df_2(b)}{db} = &-ap(\theta)[p^c(\theta_0+\delta)-p^c(\theta_0-\delta)]-a[p^1(\theta_0+\delta)-p^1(\theta_0-\delta)]\\ &-(1-c)ap^1(\theta)q^1(\theta)\log\left[\frac{p(\theta_0+\delta)q(\theta_0-\delta)}{q(\theta_0+\delta)p(\theta-\delta)}\right].\end{aligned} \tag{A.28}$$

To find the minimum/maximum of the log-likelihood ratio, set Equation (A.28) equal to 0 and solve for b. Unfortunately, simplifying the resulting equation is impossible due to the lower asymptote of $p(\theta)$ and $p(\theta_0+\delta)$. Therefore, assume $c=0$, noting that $c>0$ would result in a different b that minimizes/maximizes Equation (A.28). Setting $\frac{df_2(b)}{db}=0$, $c=0$, and dividing out the constant $-a$ yields

$$\begin{aligned}0 = p(\theta)[1-1]-&\left[\frac{\exp[a(\theta_0+\delta-b)]}{1+\exp[a(\theta_0+\delta-b)]}-\frac{\exp[a(\theta_0-\delta-b)]}{1+\exp[a(\theta_0-\delta-b)]}\right]\\ &(1-0)p^1(\theta)q^1(\theta)\log\left[\frac{\left(\frac{\exp[a(\theta_0+\delta-b)]}{1+\exp[a(\theta_0+\delta-b)]}\right)\left(\frac{1}{1+\exp[a(\theta_0-\delta-b)]}\right)}{\left(\frac{\exp[a(\theta_0-\delta-b)]}{1+\exp[a(\theta_0-\delta-b)]}\right)\left(\frac{1}{1+\exp[a(\theta_0+\delta-b)]}\right)}\right],\end{aligned}$$

$$\begin{aligned}0 = -&\left[\frac{\exp[a(\theta_0+\delta-b)]-\exp[a(\theta_0-\delta-b)]}{1+\exp[a(\theta_0+\delta-b)]+\exp[a(\theta_0-\delta-b)]+\exp[2a(\theta_0-b)]}\right]\\ &+p^1(\theta)q^1(\theta)\log\left[\frac{\exp[a(\theta_0+\delta-b)]}{\exp[a(\theta_0-\delta-b)]}\right],\end{aligned}$$

$$0 = -\left[\frac{\exp[a(\theta_0+\delta-b)]-\exp[a(\theta_0-\delta-b)]}{1+\exp[a(\theta_0+\delta-b)]+\exp[a(\theta_0-\delta-b)]+\exp[2a(\theta_0-b)]}\right]+\frac{2a\delta\exp[a(\theta-b)]}{\big(1+\exp[a(\theta-b)]\big)^2},$$

so that

$$\frac{\exp[a(\theta_0+\delta-b)]-\exp[a(\theta_0-\delta-b)]}{1+\exp[a(\theta_0+\delta-b)]+\exp[a(\theta_0-\delta-b)]+\exp[2a(\theta_0-b)]}=\frac{2a\delta\exp[a(\theta-b)]}{\big(1+\exp[a(\theta-b)]\big)^2}. \quad \text{(A.29)}$$

Next, we should cross-multiply denominators. After some manipulation, the left-hand numerator multiplied by the right-hand denominator simplifies to

$$\begin{aligned}\big(\exp[a(\theta_0+\delta-b)]-\exp[a(\theta_0-\delta-b)]\big)\big(1+\exp[a(\theta-b)]\big)^2 = \\ \exp[-1ab]\big\{\exp[a(\theta_0+\delta)]-\exp[a(\theta_0-\delta)]\big\} \\ +\exp[-2ab]\big\{2\exp[a(\theta_0+\theta+\delta)]-2\exp[a(\theta_0+\theta-\delta)]\big\} \\ +\exp[-3ab]\big\{\exp[a(\theta_0+2\theta+\delta)]-\exp[a(\theta_0+2\theta-\delta)]\big\},\end{aligned} \quad \text{(A.30)}$$

whereas the right-hand numerator multiplied by the left-hand denominator simplifies to

$$\begin{aligned}2a\delta\exp[a(\theta-b)]\big(1+\exp[a(\theta_0+\delta-b)]+\exp[a(\theta_0-\delta-b)]+\exp[2a(\theta_0-b)]\big) = \\ \exp[-1ab]\big\{2a\delta\exp[a\theta]\big\} \\ +\exp[-2ab]\big\{2a\delta\exp[a(\theta_0+\theta+\delta)]+2a\delta\exp[a(\theta_0+\theta-\delta)]\big\} \\ +\exp[-3ab]\big\{2a\delta\exp[a(2\theta_0+\theta)]\big\}.\end{aligned} \quad \text{(A.31)}$$

Subtracting Equation (A.31) from Equation (A.30) and multiplying through by $\exp[3ab]$ yields

$$\begin{aligned}
&\exp[2ab]\big\{\exp[a(\theta_0+\delta)]-\exp[a(\theta_0-\delta)]-2a\delta\exp[a\theta]\big\}\\
&\quad+\exp[ab]\big\{\exp[a(\theta_0+\theta+\delta)](2-2a\delta)-\exp[a(\theta_0+\theta-\delta)](2+2a\delta)\big\}\\
&\quad+\big\{\exp[a(\theta_0-2\theta+\delta)]-\exp[a(\theta_0-2\theta-\delta)]-2a\delta\exp[a(2\theta_0+\theta)]\big\}=0 \qquad \text{(A.32)}
\end{aligned}$$

As in the previous derivation, Equation (A.33) can be written as $x^2\omega + x\gamma + \psi = 0$, where $x = \exp[ab]$. Therefore, $x = \frac{-\gamma \pm \sqrt{\gamma^2 - 4\omega\psi}}{2\omega}$ using the quadratic formula, so that $\hat{b} = \log[x]/a$ minimizes/maximizes the expected log-likelihood ratio with respect to b.

Next, we should find each of $-\gamma$, 2ω, γ^2, and $4\omega\psi$. To simplify as much as possible, note that $2\sinh(x) = \exp(x) - \exp(-x)$ and $2\cosh(x) = \exp(x) + \exp(-x)$. First,

$$\begin{aligned}
-\gamma &= \exp[a(\theta_0+\theta-\delta)](2+2a\delta)-\exp[a(\theta_0+\theta+\delta)](2-2a\delta)\\
&= 4a\delta\cosh[a\delta]\exp[a(\theta_0+\theta)]-4\sinh[a\delta]\exp[a(\theta_0+\theta)]\\
&= 4(a\delta\cosh[a\delta]-\sinh[a\delta])\exp[a(\theta_0+\theta)]. \qquad \text{(A.33)}
\end{aligned}$$

Second,

$$\begin{aligned}
2\omega &= 2\big(\exp[a(\theta_0+\delta)]-\exp[a(\theta_0-\delta)]-2a\delta\exp[a\theta]\big)\\
&= 4\sinh[a\delta]\exp[a\theta_0]-4a\delta\exp[a\theta]. \qquad \text{(A.34)}
\end{aligned}$$

Third,

$$\begin{aligned}\gamma^2 &= \big[4(a\delta\cosh[a\delta]-\sinh[a\delta])\exp[a(\theta_0+\theta)]\big]^2\\ &= 16\big[(a\delta)^2\cosh^2[a\delta]+\sinh^2[a\delta]-2a\delta\sinh[a\delta]\cosh[a\delta]\big]\exp[2a(\theta_0+\theta)]\\ &= 16\big[(a\delta)^2\cosh^2[a\delta]-(a\delta)\sinh[2a\delta]+\sinh^2[a\delta]\big]\exp[2a(\theta_0+\theta)] \qquad \text{(A.35)}\end{aligned}$$

using the property that $2\sinh(x)\cosh(x)=\sinh(2x)$. Fourth,

$$\begin{aligned}4\omega\psi &= 4\big(\exp[a(\theta_0+\delta)]-\exp[a(\theta_0-\delta)]-2a\delta\exp[a\theta]\big)\\ &\qquad\times\big(\exp[a(\theta_0-2\theta+\delta)]-\exp[a(\theta_0-2\theta-\delta)]-2a\delta\exp[a(2\theta_0+\theta)]\big)\end{aligned}$$

$$\begin{aligned}&= 4\Big(\exp[2a(\theta_0+\theta+\delta)]-\exp[2a(\theta_0+\theta)]-2a\delta\exp[a(3\theta_0+\theta+\delta)]\\ &\qquad+\exp[2a(\theta_0+\theta-\delta)]-\exp[2a(\theta_0+\theta)]+2a\delta[a(3\theta_0+\theta-\delta)]\\ &\quad-2a\delta\exp[a(\theta_0+3\theta+\delta)]+2a\delta\exp[a(\theta_0+3\theta-\delta)]+4(a\delta)^2\exp[2a(\theta_0+\theta)]\Big)\end{aligned}$$

$$\begin{aligned}&= 16(a\delta)^2\big(\exp[2a(\theta_0+\theta)]\big)\\ &\qquad-16(a\delta)\sinh[a\delta]\big(\exp[a(3\theta_0+\theta)]+\exp[a(\theta_0+3\theta)]\\ &\qquad\qquad+8(\cosh[2a\delta]-1)\big(\exp[2a(\theta_0+\theta)]\big). \qquad \text{(A.36)}\end{aligned}$$

Finally, subtract Equation (A.36) from Equation (A.35). Note that $\sinh^2(x)=\frac{\cosh(2x)-1}{2}$, so that the third term in Equations (A.35) and (A.36) cancel, and we are left with

$$\gamma^2 - 4\omega\psi = 16(a\delta)^2\big\{(\cosh^2[a\delta] - 1)\exp[2a(\theta_0 + \theta)]\big\}$$
$$- 16(a\delta)\big\{\sinh(2a\delta)\exp[2a(\theta_0 + \theta)] - \sinh[a\delta]\big(\exp[a(3\theta_0 + \theta)] + \exp[a(\theta_0 + 3\theta)]\big)\big\}. \quad \text{(A.37)}$$

Finding the difficulty parameter that maximizes the expected log-likelihood (Equation A.25) requires square-rooting Equation (A.37), adding and subtracting that square-root from Equation (A.33), dividing the whole thing by Equation (A.34), taking the natural log of the resulting computation, and dividing that natural log by a. Unfortunately, we are yet again at an impass. Simplifying Equation (A.37) so that the square-root can be analytically determined is difficult if not intractable. However, one generally picks δ to be a *small* constant. If $\delta = 0$, then Equations (A.33), (A.34), and (A.37) evaluate to 0, so that the actual value of $\hat{b}$ at $\delta = 0$ is undefined. We can instead find $\lim_{\delta \to 0^+} \hat{b}$, where $\hat{b} = \log\left[\frac{-\gamma \pm \sqrt{\gamma^2 - 4\omega\psi}}{2\omega}\right]/a$. To start this derivation, note that if x is small, then $\sinh(x) = \frac{\exp(x) - \exp(-x)}{2} = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots \approx x$ and $\cosh(x) = \frac{\exp(x) + \exp(-x)}{2} = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots \approx 1$. Therefore,

$$\begin{aligned}
\lim_{\delta \to 0^+}(-\gamma) &= \lim_{\delta \to 0^+} 4(a\delta\cosh[a\delta] - \sinh[a\delta])\exp[a(\theta_0 + \theta)] \\
&= \lim_{\delta \to 0^+} 4(a\delta - a\delta)\exp[a(\theta_0 + \theta)] \\
&= 0. \qquad \text{(A.38)}
\end{aligned}$$

Next

$$\begin{aligned}\lim_{\delta\to 0^+}(2\omega) &= \lim_{\delta\to 0^+}\big(4\sinh[a\delta]\exp[a\theta_0] - 4a\delta\exp[a\theta]\big)\\ &= \lim_{\delta\to 0^+}\big(4a\delta\exp[a\theta_0] - 4a\delta\exp[a\theta]\big)\\ &= \lim_{\delta\to 0^+} 4a\delta(\exp[a\theta_0] - \exp[a\theta]). \qquad \text{(A.39)}\end{aligned}$$

The final piece of the puzzle is $\lim_{\delta\to 0^+}(\gamma^2 - 4\omega\psi)$. We can break down this limit into two parts. The limit of the left side of Equation (A.37) as $\delta \to 0^+$ is

$$\begin{aligned}\lim_{\delta\to 0^+} 16(a\delta)^2\big\{(\cosh^2[a\delta] - 1)\exp[2a(\theta_0+\theta)]\big\} &= \lim_{\delta\to 0^+} 16(a\delta)^2\big\{(1-1)\exp[2a(\theta_0+\theta)]\big\}\\ &= 0. \qquad \text{(A.40)}\end{aligned}$$

And the limit of the right side of Equation (A.37) as $\delta \to 0^+$ is

$$\begin{aligned}&\lim_{\delta\to 0^+} 16(a\delta)\big\{\sinh(2a\delta)\exp[2a(\theta_0+\theta)] - \sinh[a\delta]\big(\exp[a(3\theta_0+\theta)] + \exp[a(\theta_0+3\theta)]\big)\big\}\\ &\quad= \lim_{\delta\to 0^+} 16(a\delta)\big\{2a\delta\exp[2a\theta_0+\theta)] - a\delta\exp[a(3\theta_0+\theta)] - a\delta\exp[a(\theta_0+3\theta)]\big\}\\ &\quad= \lim_{\delta\to 0^+}\Big[-16(a\delta)^2\big(\exp[a(1.5\theta_0+.5\theta)] - \exp[a(.5\theta_0+1.5\theta)]\big)^2\Big] \qquad \text{(A.41)}\end{aligned}$$

Therefore, the limit of Equation (A.37) as $\delta \to 0^+$ is

$$\begin{aligned}\lim_{\delta\to 0^+}(\gamma^2 - 4\omega\psi) &= 0 - \lim_{\delta\to 0^+}\Big[-16(a\delta)^2\big(\exp[a(1.5\theta_0+.5\theta)] - \exp[a(.5\theta_0+1.5\theta)]\big)^2\Big]\\ &= \lim_{\delta\to 0^+} 16(a\delta)^2\big(\exp[a(1.5\theta_0+.5\theta)] - \exp[a(.5\theta_0+1.5\theta)]\big)^2. \qquad \text{(A.42)}\end{aligned}$$

As long as $\theta_0 \neq \theta$, the square-root function at $16(a\delta)^2\big(\exp[a(1.5\theta_0+.5\theta)]-\exp[a(.5\theta_0+1.5\theta)]\big)^2$ is continuous with a two-sided limit. Therefore, $\lim_{\delta\to 0^+}\sqrt{\gamma^2-4\omega\psi}=\sqrt{\lim_{\delta\to 0^+}(\gamma^2-4\omega\psi)}$, so that

$$\begin{aligned}
\lim_{\delta\to 0^+}\frac{-\gamma\pm\sqrt{\gamma^2-4\omega\psi}}{2\omega} &= \lim_{\delta\to 0^+}\frac{0\pm\sqrt{16(a\delta)^2\big(\exp[a(1.5\theta_0+.5\theta)]-\exp[a(.5\theta_0+1.5\theta)]\big)^2}}{4a\delta(\exp[a\theta_0]-\exp[a\theta])} \\
&= \lim_{\delta\to 0^+}\frac{\pm 4a\delta\big(\exp[a(1.5\theta_0+.5\theta)]-\exp[a(.5\theta_0+1.5\theta)]\big)}{4a\delta(\exp[a\theta_0]-\exp[a\theta])} \\
&= \frac{\pm(\exp[a\theta_0]-\exp[a\theta])\exp\left[a\left(\frac{\theta_0+\theta}{2}\right)\right]}{(\exp[a\theta_0]-\exp[a\theta])} \\
&= \pm\exp\left[a\left(\frac{\theta_0+\theta}{2}\right)\right] \qquad \text{(A.43)}
\end{aligned}$$

Finally, $\exp[x]>0$, so that $\log\big(-\exp[x]\big)$ is undefined, but $\log\big(\exp[x]\big)$ is well defined, continuous, and has a limit, so that

$$\begin{aligned}
\lim_{\delta\to 0^+}\hat{b} &= \lim_{\delta\to 0^+}\log\left[\frac{-\gamma\pm\sqrt{\gamma^2-4\omega\psi}}{2\omega}\right]/a \\
&= \log\left[\lim_{\delta\to 0^+}\left(\frac{-\gamma\pm\sqrt{\gamma^2-4\omega\psi}}{2\omega}\right)\right]/a \\
&= \log\left(\exp\left[a\left(\frac{\theta_0+\theta}{2}\right)\right]\right)/a \\
&= \frac{\theta_0+\theta}{2}. \qquad \text{(A.44)}
\end{aligned}$$

Therefore, items yielding optimal, expected log-likelihood ratios (for small δ) have difficulty parameters, b, midway between true ability, θ, and the classification bound, θ_0.

Appendix B

Tables: Aggregate over Distribution

The following tables indicate the group means, and the main effects, interactions, and corresponding effect sizes from ANOVAs predicting mean test length, classification accuracy, and various loss functions from stopping rule, item selection algorithm, correlation between dimensions, item bank, and classification bound function. Note that loss is defined as the average of Loss $= P \times I_W + J$, where I_W is an indicator function for incorrect classification, J is the number of items given to an examinee, and P is the penalty accrued for an incorrect decision.

B.1 Group Means

Table B.1: The average percentage classified correctly, number of items administered, and various loss values aggregated within each item selection algorithm assuming a compensatory classification bound function.

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
D-FI	35.2	.936	41.6	67.1	99.0
L-FI	32.8	.938	39.1	63.9	94.9
L-ELR	32.5	.937	38.8	64.0	95.6
L-KL	33.4	.937	39.7	64.9	96.4
S-KL	31.8	.940	37.8	61.8	91.8

Table B.2: The average percentage classified correctly, number of items administered, and various loss values aggregated within each item selection algorithm assuming a non-compensatory classification bound function.

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
D-FI	43.5	.917	51.8	85.0	126.6
L-FI	36.0	.912	44.8	80.0	124.0
L-ELR	35.7	.902	45.5	84.9	134.1
L-KL	37.3	.903	47.0	85.7	134.1
S-KL	39.4	.911	48.2	83.7	128.0

Table B.3: The average percentage classified correctly, number of items administered, and various loss values aggregated within each stopping rule assuming a compensatory classification bound function.

Stp. Rule		Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
P-SPRT:	$\delta = .15$	45.0	.946	50.4	72.2	99.5
P-SPRT:	$\delta = .25$	29.5	.938	35.7	60.4	91.3
C-SPRT:	$\delta = .15$	45.5	.947	50.8	72.1	98.6
C-SPRT:	$\delta = .25$	30.2	.940	36.2	60.1	90.1
M-SCSPRT:	$\delta = .15$	44.7	.947	50.0	71.1	95.3
M-SCSPRT:	$\delta = .25$	30.1	.941	36.0	59.8	89.5
M-GLR:	$\delta = .15$	29.7	.934	36.2	62.5	95.3
M-GLR:	$\delta = .25$	23.9	.926	31.3	61.1	98.3
BCR:	$\alpha = .05$	30.7	.937	36.9	62.0	93.4
BCR:	$\alpha = .10$	22.4	.920	30.4	62.2	102.0

Table B.4: The average percentage classified correctly, number of items administered, and various loss values aggregated within each stopping rule assuming a non-compensatory classification bound function.

Stp. Rule		Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
P-SPRT:	$\delta = .15$	40.9	.896	51.3	93.0	145.0
P-SPRT:	$\delta = .25$	25.8	.881	37.7	85.2	144.6
C-SPRT:	$\delta = .15$	55.9	.923	63.6	94.2	132.6
C-SPRT:	$\delta = .25$	40.0	.917	48.2	81.3	122.6
M-SCSPRT:	$\delta = .15$	49.1	.905	58.6	96.5	143.8
M-SCSPRT:	$\delta = .25$	36.4	.902	46.3	85.5	134.6
M-GLR:	$\delta = .15$	39.2	.917	47.5	80.5	121.8
M-GLR:	$\delta = .25$	36.3	.915	44.8	78.9	121.6
BCR:	$\alpha = .05$	34.3	.920	42.3	74.1	114.0
BCR:	$\alpha = .10$	25.8	.913	34.5	69.4	113.0

Table B.5: The average percentage classified correctly, number of items administered, and various loss values aggregated within each item bank by dimension correlation assuming a compensatory classification bound function.

Between Multidimensional Item Bank

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
$\rho = .00$	40.5	.925	48.0	78.1	115.7
$\rho = .33$	37.8	.932	44.6	71.9	106.0
$\rho = .67$	35.5	.938	41.7	66.4	97.2

Within Multidimensional Item Bank

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
$\rho = .00$	30.3	.937	36.6	61.7	93.1
$\rho = .33$	28.2	.944	33.9	56.3	84.5
$\rho = .67$	26.6	.950	31.6	51.7	76.9

Table B.6: The average percentage classified correctly, number of items administered, and various loss values aggregated within each item bank by dimension correlation assuming a non-compensatory classification bound function.

Between Multidimensional Item Bank

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
$\rho = .00$	32.8	.906	42.2	79.7	126.6
$\rho = .33$	32.6	.913	41.3	76.1	119.6
$\rho = .67$	32.9	.916	41.4	75.1	117.3

Within Multidimensional Item Bank

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
$\rho = .00$	44.7	.904	54.3	92.8	141.0
$\rho = .33$	44.0	.908	53.2	90.2	135.4
$\rho = .67$	43.2	.908	52.4	89.2	135.3

Table B.7: The average percentage classified correctly, number of items administered, and various loss values aggregated within each dimension correlation by item selection algorithm assuming a compensatory classification bound function.

$\text{Cor}(\theta_1, \theta_2) = .00$

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
D-FI	37.7	.929	44.8	73.2	108.7
L-FI	35.0	.931	41.8	69.2	103.5
L-ELR	34.8	.931	41.7	69.2	103.7
L-KL	35.5	.930	42.5	70.4	105.3
S-KL	34.0	.933	40.6	67.4	100.8

$\text{Cor}(\theta_1, \theta_2) = .33$

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
D-FI	35.1	.936	41.5	67.1	99.1
L-FI	32.7	.938	38.9	63.8	94.9
L-ELR	32.3	.937	38.6	64.0	95.8
L-KL	33.3	.938	39.4	64.1	94.9
S-KL	31.7	.940	37.7	61.5	91.3

$\text{Cor}(\theta_1, \theta_2) = .67$

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
D-FI	32.7	.943	38.3	61.0	89.3
L-FI	30.9	.944	36.4	58.6	86.4
L-ELR	30.4	.943	36.1	58.9	87.3
L-KL	31.4	.942	37.1	60.2	89.0
S-KL	29.8	.947	35.2	56.6	83.3

Table B.8: The average percentage classified correctly, number of items administered, and various loss values aggregated within each dimension correlation by item selection algorithm assuming a non-compensatory classification bound function.

$\text{Cor}(\theta_1, \theta_2) = .00$

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
D-FI	44.2	.915	52.7	86.6	129.0
L-FI	36.4	.908	45.6	82.5	128.6
L-ELR	36.0	.896	46.4	88.0	140.0
L-KL	37.8	.898	48.0	89.0	140.2
S-KL	39.4	.908	48.6	85.2	131.1

$\text{Cor}(\theta_1, \theta_2) = .33$

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
D-FI	43.6	.919	51.8	84.3	124.9
L-FI	35.8	.913	44.5	79.3	122.7
L-ELR	35.6	.903	45.3	84.3	133.0
L-KL	37.3	.904	46.8	85.0	132.9
S-KL	39.2	.913	48.0	82.9	126.5

$\text{Cor}(\theta_1, \theta_2) = .67$

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
D-FI	42.8	.917	51.1	84.3	125.8
L-FI	35.8	.915	44.3	78.2	120.6
L-ELR	35.3	.906	44.7	82.4	129.4
L-KL	37.0	.908	46.2	83.1	129.1
S-KL	39.5	.913	48.2	83.0	126.5

Table B.9: The average percentage classified correctly, number of items administered, and various loss values aggregated within each dimension correlation by stopping rule assuming a compensatory classification bound function.

$\text{Cor}(\theta_1, \theta_2) = .00$

Stp. Rule		Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
P-SPRT:	$\delta = .15$	47.8	.940	53.8	77.9	108.1
P-SPRT:	$\delta = .25$	31.3	.932	38.1	65.4	99.6
C-SPRT:	$\delta = .15$	48.2	.941	54.1	77.5	106.8
C-SPRT:	$\delta = .25$	32.0	.933	38.6	65.3	98.6
M-SCSPRT:	$\delta = .15$	47.4	.942	53.2	76.3	105.1
M-SCSPRT:	$\delta = .25$	31.8	.932	38.6	65.9	99.9
M-GLR:	$\delta = .15$	31.8	.927	39.1	68.3	104.7
M-GLR:	$\delta = .25$	25.4	.918	33.6	66.3	107.2
BCR:	$\alpha = .05$	33.7	.931	40.7	68.4	103.1
BCR:	$\alpha = .10$	24.5	.914	33.1	67.7	110.8

$\text{Cor}(\theta_1, \theta_2) = .33$

Stp. Rule		Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
P-SPRT:	$\delta = .15$	44.9	.945	50.4	72.3	99.7
P-SPRT:	$\delta = .25$	29.4	.939	35.5	60.0	90.7
C-SPRT:	$\delta = .15$	45.4	.947	50.7	72.1	98.8
C-SPRT:	$\delta = .25$	30.0	.940	36.0	59.9	89.8
M-SCSPRT:	$\delta = .15$	44.5	.946	49.9	71.4	98.3
M-SCSPRT:	$\delta = .25$	39.9	.943	35.6	58.5	87.1
M-GLR:	$\delta = .15$	29.6	.935	36.1	36.1	94.4
M-GLR:	$\delta = .25$	23.8	.925	31.3	31.3	98.6
BCR:	$\alpha = .05$	30.3	.937	36.6	61.6	92.9
BCR:	$\alpha = .10$	22.3	.920	30.2	62.1	101.9

$\text{Cor}(\theta_1, \theta_2) = .67$

Stp. Rule		Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
P-SPRT:	$\delta = .15$	42.2	.952	47.0	66.4	90.6
P-SPRT:	$\delta = .25$	27.8	.944	33.4	55.7	83.7
C-SPRT:	$\delta = .15$	42.9	.953	47.7	47.7	90.1
C-SPRT:	$\delta = .25$	28.5	.946	33.8	33.8	82.0
M-SCSPRT:	$\delta = .15$	42.1	.953	46.8	46.8	89.3
M-SCSPRT:	$\delta = .25$	28.4	.947	33.7	33.7	81.6
M-GLR:	$\delta = .15$	27.6	.941	33.5	33.5	86.7
M-GLR:	$\delta = .25$	22.4	.933	29.1	29.1	89.1
BCR:	$\alpha = .05$	27.9	.944	33.6	33.6	84.3
BCR:	$\alpha = .10$	20.5	.927	27.8	27.8	93.3

Table B.10: The average percentage classified correctly, number of items administered, and various loss values aggregated within each dimension correlation by stopping rule assuming a non-compensatory classification bound function.

$\text{Cor}(\theta_1, \theta_2) = .00$

Stp. Rule		Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
P-SPRT:	$\delta = .15$	41.3	.886	52.7	98.2	155.1
P-SPRT:	$\delta = .25$	26.3	.872	39.1	90.5	154.7
C-SPRT:	$\delta = .15$	56.0	.922	63.8	95.2	134.5
C-SPRT:	$\delta = .25$	39.9	.914	48.5	82.8	125.6
M-SCSPRT:	$\delta = .15$	49.6	.902	59.4	98.6	147.6
M-SCSPRT:	$\delta = .25$	36.5	.900	46.5	86.4	136.2
M-GLR:	$\delta = .15$	39.3	.914	47.9	82.2	125.1
M-GLR:	$\delta = .25$	36.5	.909	45.5	81.8	127.2
BCR:	$\alpha = .05$	35.5	.918	43.6	76.3	117.1
BCR:	$\alpha = .10$	26.8	.912	35.6	70.7	114.6

$\text{Cor}(\theta_1, \theta_2) = .33$

Stp. Rule		Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
P-SPRT:	$\delta = .15$	41.0	.898	51.2	92.2	143.3
P-SPRT:	$\delta = .25$	25.9	.884	37.5	84.0	142.1
C-SPRT:	$\delta = .15$	55.7	.925	63.2	63.2	131.1
C-SPRT:	$\delta = .25$	39.8	.918	48.0	48.0	122.1
M-SCSPRT:	$\delta = .15$	48.9	.908	58.1	58.1	141.2
M-SCSPRT:	$\delta = .25$	36.2	.902	46.1	46.1	134.7
M-GLR:	$\delta = .15$	38.9	.920	46.9	46.9	119.3
M-GLR:	$\delta = .25$	36.0	.917	44.3	44.3	119.1
BCR:	$\alpha = .05$	34.6	.920	42.5	42.5	114.1
BCR:	$\alpha = .10$	26.0	.913	34.7	34.7	113.1

$\text{Cor}(\theta_1, \theta_2) = .67$

Stp. Rule		Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
P-SPRT:	$\delta = .15$	40.5	.904	50.1	88.5	136.6
P-SPRT:	$\delta = .25$	25.2	.888	36.4	81.2	137.1
C-SPRT:	$\delta = .15$	56.0	.924	63.6	94.1	132.2
C-SPRT:	$\delta = .25$	40.2	.920	48.2	80.2	120.2
M-SCSPRT:	$\delta = .15$	49.0	.906	58.3	95.8	142.6
M-SCSPRT:	$\delta = .25$	36.6	.904	46.2	84.8	132.9
M-GLR:	$\delta = .15$	39.4	.918	47.6	80.3	121.1
M-GLR:	$\delta = .25$	36.3	.918	44.5	77.4	118.4
BCR:	$\alpha = .05$	32.9	.922	40.7	71.8	110.7
BCR:	$\alpha = .10$	24.7	.913	33.3	67.9	111.2

Table B.11: The average percentage classified correctly, number of items administered, and various loss values aggregated within each item bank by item selection algorithm assuming a compensatory classification bound function.

Between Multidimensional Item Bank

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
D-FI	38.3	.932	45.1	72.1	105.8
L-FI	38.2	.932	45.0	72.2	106.3
L-ELR	37.6	.930	44.7	72.8	108.0
L-KL	38.8	.931	45.7	73.5	108.3
S-KL	36.8	.934	43.4	70.0	103.1

Within Multidimensional Item Bank

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
D-FI	32.0	.940	38.1	62.1	92.3
L-FI	27.5	.944	33.1	55.6	83.6
L-ELR	27.4	.944	33.0	55.3	83.2
L-KL	28.0	.944	33.7	56.3	84.5
S-KL	26.9	.946	32.2	53.7	80.5

Table B.12: The average percentage classified correctly, number of items administered, and various loss values aggregated within each item bank by item selection algorithm assuming a non-compensatory classification bound function.

Between Multidimensional Item Bank

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
D-FI	40.3	.923	48.1	79.0	117.8
L-FI	30.9	.910	39.9	76.0	121.1
L-ELR	29.5	.896	39.9	81.6	133.6
L-KL	30.8	.909	39.9	76.1	121.5
S-KL	32.5	.921	40.4	72.2	111.9

Within Multidimensional Item Bank

Select	Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
D-FI	46.8	.911	55.6	91.1	135.4
L-FI	41.1	.914	49.7	84.0	126.9
L-ELR	41.8	.907	51.1	88.2	134.7
L-KL	43.9	.897	54.2	95.3	146.6
S-KL	46.3	.902	56.1	95.2	144.1

Table B.13: The average percentage classified correctly, number of items administered, and various loss values aggregated within each item bank by stopping rule assuming a compensatory classification bound function.

Between Multidimensional Item Bank

Stp. Rule		Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
P-SPRT:	$\delta = .15$	52.4	.939	58.5	83.0	113.5
P-SPRT:	$\delta = .25$	35.2	.935	41.7	67.6	100.1
C-SPRT:	$\delta = .15$	52.5	.942	58.3	81.7	110.8
C-SPRT:	$\delta = .25$	35.4	.937	41.8	67.1	98.7
M-SCSPRT:	$\delta = .15$	51.2	.942	57.0	80.4	109.6
M-SCSPRT:	$\delta = .25$	35.3	.936	41.7	67.1	98.9
M-GLR:	$\delta = .15$	32.6	.926	40.0	69.6	106.6
M-GLR:	$\delta = .25$	26.9	.918	35.1	68.1	109.3
BCR:	$\alpha = .05$	33.6	.931	40.5	68.2	102.8
BCR:	$\alpha = .10$	24.4	.912	33.2	68.5	112.5

Within Multidimensional Item Bank

Stp. Rule		Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
P-SPRT:	$\delta = .15$	37.5	.952	42.3	61.4	85.4
P-SPRT:	$\delta = .25$	23.8	.941	29.7	52.4	82.5
C-SPRT:	$\delta = .15$	38.6	.952	43.3	62.4	86.3
C-SPRT:	$\delta = .25$	24.9	.943	30.6	53.2	81.6
M-SCSPRT:	$\delta = .15$	38.2	.953	42.9	61.9	85.6
M-SCSPRT:	$\delta = .25$	24.8	.945	30.3	52.4	80.1
M-GLR:	$\delta = .15$	26.7	.943	32.5	55.3	83.9
M-GLR:	$\delta = .25$	20.9	.934	27.5	54.1	87.3
BCR:	$\alpha = .05$	27.7	.944	33.4	55.9	84.1
BCR:	$\alpha = .10$	20.5	.929	27.6	56.0	91.5

Table B.14: The average percentage classified correctly, number of items administered, and various loss values aggregated within each item bank by stopping rule assuming a non-compensatory classification bound function.

Between Multidimensional Item Bank

Stp. Rule		Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
P-SPRT:	$\delta = .15$	41.6	.892	52.4	95.7	149.8
P-SPRT:	$\delta = .25$	27.1	.883	38.8	85.7	144.2
C-SPRT:	$\delta = .15$	45.6	.934	52.1	78.4	111.3
C-SPRT:	$\delta = .25$	29.9	.923	37.6	68.3	106.8
M-SCSPRT:	$\delta = .15$	35.7	.898	45.9	86.7	137.7
M-SCSPRT:	$\delta = .25$	24.1	.892	34.9	78.1	132.1
M-GLR:	$\delta = .15$	34.4	.926	41.8	71.4	108.4
M-GLR:	$\delta = .25$	31.8	.922	39.7	71.0	110.2
BCR:	$\alpha = .05$	33.1	.929	40.2	68.7	104.4
BCR:	$\alpha = .10$	24.7	.918	32.9	65.8	106.9

Within Multidimensional Item Bank

Stp. Rule		Test Len.	Class Acc.	Loss ($P = 100$)	Loss ($P = 500$)	Loss($P = 1000$)
P-SPRT:	$\delta = .15$	40.3	.900	50.2	90.2	140.2
P-SPRT:	$\delta = .25$	24.5	.890	36.6	84.8	145.0
C-SPRT:	$\delta = .15$	66.2	.912	75.0	110.1	153.9
C-SPRT:	$\delta = .25$	50.0	.912	58.9	94.3	138.5
M-SCSPRT:	$\delta = .15$	62.5	.913	71.3	106.2	149.9
M-SCSPRT:	$\delta = .25$	48.8	.912	57.6	93.0	137.1
M-GLR:	$\delta = .15$	44.1	.909	53.2	89.7	135.3
M-GLR:	$\delta = .25$	40.7	.908	49.9	86.8	132.9
BCR:	$\alpha = .05$	35.5	.912	44.3	79.6	123.6
BCR:	$\alpha = .10$	27.0	.908	36.2	73.0	119.0

Table B.15: The average percentage classified correctly and number of items administered within each item selection algorithm by stopping rule assuming a compensatory classification bound function and a between multidimensional item bank.

Average Test Length

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	52.8	52.8	51.8	53.1	51.5
P-SPRT:	$\delta = .25$	35.6	35.7	34.6	35.6	34.4
C-SPRT:	$\delta = .15$	53.2	52.8	51.3	53.9	51.3
C-SPRT:	$\delta = .25$	35.6	35.8	35.0	36.5	34.2
M-SCSPRT:	$\delta = .15$	51.7	51.4	50.2	52.5	50.1
M-SCSPRT:	$\delta = .25$	35.5	35.4	35.1	36.5	34.1
M-GLR:	$\delta = .15$	33.3	33.3	31.9	33.3	31.1
M-GLR:	$\delta = .25$	27.3	27.2	26.9	27.7	25.4
BCR:	$\alpha = .05$	33.6	33.4	34.3	34.3	32.4
BCR:	$\alpha = .10$	24.4	24.2	25.2	24.4	23.7

Classification Accuracy

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	.938	.941	.938	.936	.941
P-SPRT:	$\delta = .25$	.938	.935	.931	.934	.938
C-SPRT:	$\delta = .15$	.939	.941	.942	.941	.944
C-SPRT:	$\delta = .25$	.936	.935	.935	.938	.940
M-SCSPRT:	$\delta = .15$	.942	.941	.940	.943	.942
M-SCSPRT:	$\delta = .25$	.939	.935	.936	.934	.938
M-GLR:	$\delta = .15$	.932	.927	.920	.923	.928
M-GLR:	$\delta = .25$	.919	.922	.911	.916	.920
BCR:	$\alpha = .05$	.932	.931	.929	.929	.932
BCR:	$\alpha = .10$	.909	.911	.914	.912	.913

Table B.16: Various loss values within each item selection algorithm by stopping rule assuming a compensatory classification bound function and a between multidimensional item bank.

Loss ($P = 100$)

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	59.0	58.7	58.0	59.5	57.4
P-SPRT:	$\delta = .25$	41.8	42.2	41.5	42.2	40.6
C-SPRT:	$\delta = .15$	59.2	58.7	57.0	59.8	56.9
C-SPRT:	$\delta = .25$	42.0	42.4	41.5	42.7	40.3
M-SCSPRT:	$\delta = .15$	57.5	57.4	56.1	58.3	55.9
M-SCSPRT:	$\delta = .25$	41.6	42.0	41.5	43.1	40.3
M-GLR:	$\delta = .15$	40.2	40.5	39.9	41.0	38.3
M-GLR:	$\delta = .25$	35.4	35.0	35.7	36.2	33.4
BCR:	$\alpha = .05$	40.4	40.2	41.4	41.4	39.1
BCR:	$\alpha = .10$	33.6	33.1	33.8	33.1	32.3

Loss ($P = 500$)

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	83.7	82.1	82.9	85.1	81.1
P-SPRT:	$\delta = .25$	66.7	68.1	69.3	68.7	65.4
C-SPRT:	$\delta = .15$	83.5	82.2	80.2	83.3	79.2
C-SPRT:	$\delta = .25$	67.4	68.2	67.5	67.7	64.4
M-SCSPRT:	$\delta = .15$	80.5	81.1	79.9	81.2	79.0
M-SCSPRT:	$\delta = .25$	66.0	68.0	67.0	69.6	65.0
M-GLR:	$\delta = .15$	67.6	69.6	71.9	71.8	67.1
M-GLR:	$\delta = .25$	67.7	66.4	71.1	69.9	65.3
BCR:	$\alpha = .05$	67.5	67.7	69.9	69.7	66.1
BCR:	$\alpha = .10$	70.1	68.8	68.2	68.2	67.0

Loss ($P = 1000$)

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	114.5	111.4	114.0	117.1	110.7
P-SPRT:	$\delta = .25$	97.9	100.5	104.0	101.9	96.3
C-SPRT:	$\delta = .15$	113.8	111.5	109.1	112.6	107.0
C-SPRT:	$\delta = .25$	99.2	100.7	100.0	98.8	94.6
M-SCSPRT:	$\delta = .15$	109.4	110.8	109.7	109.9	107.9
M-SCSPRT:	$\delta = .25$	96.5	100.6	98.8	102.8	95.8
M-GLR:	$\delta = .15$	101.8	106.0	111.9	110.3	103.2
M-GLR:	$\delta = .25$	108.1	105.5	115.4	112.1	105.3
BCR:	$\alpha = .05$	101.3	102.1	105.4	105.2	99.9
BCR:	$\alpha = .10$	115.8	113.4	111.2	112.0	110.3

Table B.17: The average percentage classified correctly and number of items administered within each item selection algorithm by stopping rule assuming a compensatory classification bound function and a within multidimensional item bank.

Average Test Length

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	43.2	36.6	35.5	36.5	35.7
P-SPRT:	$\delta = .25$	28.0	23.2	22.8	22.9	22.1
C-SPRT:	$\delta = .15$	44.0	37.2	36.5	38.3	36.6
C-SPRT:	$\delta = .25$	28.9	23.7	23.8	24.6	23.4
M-SCSPRT:	$\delta = .15$	43.4	37.0	36.0	38.2	36.3
M-SCSPRT:	$\delta = .25$	28.9	23.6	23.6	24.4	23.5
M-GLR:	$\delta = .15$	29.2	26.1	26.2	27.0	25.2
M-GLR:	$\delta = .25$	23.1	20.3	20.3	20.9	19.7
BCR:	$\alpha = .05$	29.8	27.0	28.0	27.4	26.4
BCR:	$\alpha = .10$	21.7	20.0	21.0	20.0	19.7

Classification Accuracy

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	.953	.952	.952	.951	.953
P-SPRT:	$\delta = .25$	.941	.941	.937	.941	.947
C-SPRT:	$\delta = .15$	.949	.954	.952	.955	.952
C-SPRT:	$\delta = .25$	.939	.943	.946	.945	.944
M-SCSPRT:	$\delta = .15$	.951	.952	.953	.953	.955
M-SCSPRT:	$\delta = .25$	.943	.947	.944	.942	.947
M-GLR:	$\delta = .15$	.937	.944	.944	.943	.947
M-GLR:	$\delta = .25$	.926	.933	.936	.932	.940
BCR:	$\alpha = .05$	.939	.942	.947	.043	.947
BCR:	$\alpha = .10$	.920	.930	.931	.931	.932

Table B.18: Various loss values within each item selection algorithm by stopping rule assuming a compensatory classification bound function and a within multidimensional item bank.

Loss ($P = 100$)

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	47.9	41.4	40.3	41.4	40.4
P-SPRT:	$\delta = .25$	33.9	29.1	29.1	28.7	27.4
C-SPRT:	$\delta = .15$	49.1	41.8	41.4	42.9	41.5
C-SPRT:	$\delta = .25$	35.1	29.4	29.2	30.1	29.0
M-SCSPRT:	$\delta = .15$	48.4	41.8	40.7	42.9	40.8
M-SCSPRT:	$\delta = .25$	34.6	28.9	29.2	30.1	28.8
M-GLR:	$\delta = .15$	35.6	31.7	31.8	32.8	30.5
M-GLR:	$\delta = .25$	30.5	27.0	26.7	27.7	25.7
BCR:	$\alpha = .05$	35.9	32.8	33.3	33.1	31.7
BCR:	$\alpha = .10$	29.7	27.0	27.9	26.9	26.5

Loss ($P = 500$)

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	66.7	60.7	59.6	60.9	59.3
P-SPRT:	$\delta = .25$	57.5	52.9	54.4	52.2	48.8
C-SPRT:	$\delta = .15$	69.4	60.3	60.6	61.0	60.8
C-SPRT:	$\delta = .25$	59.6	52.1	50.7	52.3	51.4
M-SCSPRT:	$\delta = .15$	68.2	60.9	59.7	61.9	58.9
M-SCSPRT:	$\delta = .25$	57.4	50.1	51.6	53.3	49.8
M-GLR:	$\delta = .15$	60.9	54.0	54.2	55.8	51.8
M-GLR:	$\delta = .25$	60.0	53.6	52.4	54.8	49.6
BCR:	$\alpha = .05$	60.1	56.0	54.5	56.0	53.0
BCR:	$\alpha = .10$	61.6	54.9	55.4	54.5	53.6

Loss ($P = 1000$)

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	90.2	84.7	83.7	85.4	82.9
P-SPRT:	$\delta = .25$	87.0	82.6	85.9	81.6	75.5
C-SPRT:	$\delta = .15$	94.9	83.4	84.7	83.6	85.0
C-SPRT:	$\delta = .25$	90.3	80.5	77.6	80.0	79.5
M-SCSPRT:	$\delta = .15$	92.9	84.8	83.4	85.6	81.4
M-SCSPRT:	$\delta = .25$	85.9	76.6	79.6	82.2	76.1
M-GLR:	$\delta = .15$	92.6	81.9	82.2	84.5	78.3
M-GLR:	$\delta = .25$	96.9	86.9	84.4	88.8	79.6
BCR:	$\alpha = .05$	90.5	84.9	81.0	84.5	79.6
BCR:	$\alpha = .10$	101.5	89.8	89.7	89.0	87.5

Table B.19: The average percentage classified correctly and number of items administered within each item selection algorithm by stopping rule assuming a non-compensatory classification bound function and a between multidimensional item bank.

Average Test Length

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	55.5	35.1	39.0	35.3	43.2
P-SPRT:	$\delta = .25$	38.1	21.9	26.8	22.0	26.8
C-SPRT:	$\delta = .15$	58.9	40.4	41.9	40.5	46.0
C-SPRT:	$\delta = .25$	39.6	25.1	28.4	24.8	31.4
M-SCSPRT:	$\delta = .15$	45.2	40.0	19.8	39.7	34.0
M-SCSPRT:	$\delta = .25$	32.7	24.9	15.3	25.0	22.6
M-GLR:	$\delta = .15$	38.5	33.0	33.7	33.1	33.6
M-GLR:	$\delta = .25$	33.7	31.2	31.7	30.9	31.6
BCR:	$\alpha = .05$	35.1	32.6	33.5	32.2	32.0
BCR:	$\alpha = .10$	25.9	24.5	25.3	24.4	23.4

Classification Accuracy

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	.937	.855	.878	.856	.934
P-SPRT:	$\delta = .25$	.930	.947	.868	.849	.919
C-SPRT:	$\delta = .15$	.936	.934	.935	.931	.935
C-SPRT:	$\delta = .25$	.932	.919	.926	.917	.922
M-SCSPRT:	$\delta = .15$	.900	.931	.826	.932	.902
M-SCSPRT:	$\delta = .25$	.898	.919	.828	.919	.897
M-GLR:	$\delta = .15$	.931	.926	.929	.921	.924
M-GLR:	$\delta = .25$	.923	.924	.921	.919	.921
BCR:	$\alpha = .05$	.925	.930	.929	.932	.928
BCR:	$\alpha = .10$	.914	.915	.919	.917	.924

Table B.20: Various loss values within each item selection algorithm by stopping rule assuming a non-compensatory classification bound function and a between multidimensional item bank.

Loss ($P = 100$)

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	61.8	49.6	51.2	49.7	49.8
P-SPRT:	$\delta = .25$	45.1	37.2	40.0	37.0	34.9
C-SPRT:	$\delta = .15$	65.4	47.0	48.4	47.4	52.5
C-SPRT:	$\delta = .25$	46.5	33.2	35.8	33.1	39.3
M-SCSPRT:	$\delta = .15$	55.1	46.9	37.2	46.5	43.8
M-SCSPRT:	$\delta = .25$	42.9	33.1	32.5	33.1	32.8
M-GLR:	$\delta = .15$	45.4	40.4	40.8	41.0	41.2
M-GLR:	$\delta = .25$	41.4	38.9	39.6	39.0	39.5
BCR:	$\alpha = .05$	42.6	39.6	40.6	39.1	39.2
BCR:	$\alpha = .10$	34.5	33.0	33.4	32.6	31.0

Loss ($P = 500$)

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	87.2	107.8	99.8	107.4	76.4
P-SPRT:	$\delta = .25$	72.9	98.3	92.9	97.3	67.1
C-SPRT:	$\delta = .15$	91.1	73.4	74.3	75.1	78.4
C-SPRT:	$\delta = .25$	73.8	65.7	65.4	66.1	70.6
M-SCSPRT:	$\delta = .15$	95.0	74.7	106.9	73.8	83.2
M-SCSPRT:	$\delta = .25$	83.8	65.7	101.5	65.6	73.9
M-GLR:	$\delta = .15$	72.8	70.2	69.3	72.7	71.8
M-GLR:	$\delta = .25$	72.3	69.4	71.0	71.4	71.0
BCR:	$\alpha = .05$	72.4	67.7	68.9	66.4	68.1
BCR:	$\alpha = .10$	69.0	66.8	65.8	65.8	61.5

Loss ($P = 1000$)

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	118.9	180.5	160.5	179.4	109.6
P-SPRT:	$\delta = .25$	107.7	174.6	159.0	172.6	107.4
C-SPRT:	$\delta = .15$	123.2	106.3	106.6	109.6	110.8
C-SPRT:	$\delta = .25$	108.0	106.3	102.4	107.4	109.7
M-SCSPRT:	$\delta = .15$	144.7	109.4	194.0	107.8	132.3
M-SCSPRT:	$\delta = .25$	134.9	106.4	187.8	106.2	125.2
M-GLR:	$\delta = .15$	107.2	107.5	104.9	112.2	110.0
M-GLR:	$\delta = .25$	110.9	107.6	110.3	111.8	110.3
BCR:	$\alpha = .05$	109.7	102.8	104.3	100.7	104.3
BCR:	$\alpha = .10$	112.2	109.1	106.4	107.3	99.7

Table B.21: The average percentage classified correctly and number of items administered within each item selection algorithm by stopping rule assuming a non-compensatory classification bound function and a within multidimensional item bank.

Average Test Length

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	51.7	41.5	36.7	34.9	36.5
P-SPRT:	$\delta = .25$	31.3	23.8	23.2	21.4	22.8
C-SPRT:	$\delta = .15$	69.0	62.4	62.8	66.3	70.7
C-SPRT:	$\delta = .25$	51.2	43.7	46.2	51.4	57.6
M-SCSPRT:	$\delta = .15$	65.8	56.2	59.3	63.7	67.8
M-SCSPRT:	$\delta = .25$	49.9	42.2	44.9	50.6	56.4
M-GLR:	$\delta = .15$	44.1	41.3	42.3	45.4	47.4
M-GLR:	$\delta = .25$	40.2	37.7	38.8	42.6	44.3
BCR:	$\alpha = .05$	36.7	35.3	36.3	35.4	34.0
BCR:	$\alpha = .10$	27.6	27.3	27.4	27.4	25.3

Classification Accuracy

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	.910	.909	.893	.889	.899
P-SPRT:	$\delta = .25$	.891	.885	.866	.872	.884
C-SPRT:	$\delta = .15$	.915	.921	.921	.901	.904
C-SPRT:	$\delta = .25$	.913	.921	.917	.905	.901
M-SCSPRT:	$\delta = .15$	.917	.924	.916	.903	.903
M-SCSPRT:	$\delta = .25$	.916	.919	.915	.902	.906
M-GLR:	$\delta = .15$	.912	.918	.913	.900	.902
M-GLR:	$\delta = .25$	.913	.916	.910	.898	.902
BCR:	$\alpha = .05$	.916	.919	.912	.902	.911
BCR:	$\alpha = .10$	.910	.911	.908	.901	.910

Table B.22: Various loss values within each item selection algorithm by stopping rule assuming a non-compensatory classification bound function and a within multidimensional item bank.

Loss ($P = 100$)

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	60.6	50.6	47.4	46.0	46.6
P-SPRT:	$\delta = .25$	42.2	35.3	36.6	34.3	34.4
C-SPRT:	$\delta = .15$	77.5	70.2	70.7	76.2	80.3
C-SPRT:	$\delta = .25$	59.9	51.7	54.4	60.9	67.5
M-SCSPRT:	$\delta = .15$	74.1	63.8	67.7	73.3	77.5
M-SCSPRT:	$\delta = .25$	58.3	50.3	53.4	60.5	65.8
M-GLR:	$\delta = .15$	52.9	49.5	51.0	55.4	57.2
M-GLR:	$\delta = .25$	48.9	46.1	47.7	52.9	54.1
BCR:	$\alpha = .05$	45.1	43.4	45.1	45.2	42.9
BCR:	$\alpha = .10$	36.6	36.1	36.6	37.3	34.3

Loss ($P = 500$)

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	96.5	87.0	90.2	90.2	87.2
P-SPRT:	$\delta = .25$	85.6	81.5	90.2	85.5	80.9
C-SPRT:	$\delta = .15$	111.6	101.7	102.3	115.8	118.9
C-SPRT:	$\delta = .25$	94.5	83.4	87.5	99.0	106.9
M-SCSPRT:	$\delta = .15$	107.5	94.3	101.3	111.9	116.1
M-SCSPRT:	$\delta = .25$	91.7	82.7	87.4	99.8	103.2
M-GLR:	$\delta = .15$	88.2	82.4	86.0	95.4	96.4
M-GLR:	$\delta = .25$	83.8	79.7	83.6	93.8	93.2
BCR:	$\alpha = .05$	78.6	76.0	80.3	84.3	78.7
BCR:	$\alpha = .10$	72.6	71.6	73.3	77.1	70.3

Loss ($P = 1000$)

Stop↓	Select→	D-FI	L-FI	L-ELR	L-KL	S-KL
P-SPRT:	$\delta = .15$	141.4	132.4	143.7	145.5	137.9
P-SPRT:	$\delta = .25$	139.9	139.3	157.3	149.6	138.9
C-SPRT:	$\delta = .15$	154.1	141.0	141.9	165.3	167.1
C-SPRT:	$\delta = .25$	137.8	123.1	128.8	146.6	156.1
M-SCSPRT:	$\delta = .15$	149.2	132.3	143.4	160.2	164.5
M-SCSPRT:	$\delta = .25$	133.5	123.2	129.8	149.0	150.0
M-GLR:	$\delta = .15$	132.4	123.5	129.8	145.5	145.3
M-GLR:	$\delta = .25$	127.4	121.7	128.4	144.9	142.2
BCR:	$\alpha = .05$	120.4	116.7	124.4	133.1	123.3
BCR:	$\alpha = .10$	117.7	115.9	119.2	126.8	115.3

B.2 Effect Sizes

Table B.23: The sums of squares (Sum Sq.), $\eta^2 = \frac{\text{SSF}}{\text{SST}}$, and $\omega^2 = \frac{\text{SSF} - \text{df}_F \times \text{MSE}}{\text{SST} + \text{MSE}}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting mean classification accuracy given a compensatory classification bound function. The ANOVA was run with all main effects, two-way interactions, and three-way interactions.

Variance Type	Sum Sq.	η^2	ω^2
Correlation (Cor)	0.00841	.185	.185
Item Bank (Bank)	0.01055	.233	.232
Select Alg. (Select)	0.00054	.012	.011
Stop Rule (Stop)	0.02115	.466	.464
Cor by Bank	0.00001	.000	.000
Cor by Select	0.00004	.001	.000
Cor by Stop	0.00010	.002	.000
Bank by Select	0.00044	.010	.009
Bank by Stop	0.00111	.025	.023
Select by Stop	0.00066	.015	.007
Cor by Bank by Select	0.00006	.001	.001
Cor by Bank by Stop	0.00014	.003	.001
Cor by Select by Stop	0.00038	.008	.004
Bank by Select by Stop	0.00053	.012	.000
Residuals	0.00123		
Total	0.04535		

Table B.24: The sums of squares (Sum Sq.), $\eta^2 = \frac{\text{SSF}}{\text{SST}}$, and $\omega^2 = \frac{\text{SSF} - \text{df}_F \times \text{MSE}}{\text{SST} + \text{MSE}}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting mean test length given a compensatory classification bound function. The ANOVA was run with all main effects, two-way interactions, and three-way interactions.

Variance Type	Sum Sq.	η^2	ω^2
Correlation (Cor)	947.55	.032	.032
Item Bank (Bank)	6901.34	.230	.230
Select Alg. (Select)	382.07	.013	.013
Stop Rule (Stop)	20301.92	.676	.676
Cor by Bank	20.47	.001	.001
Cor by Select	6.07	.000	.000
Cor by Stop	49.39	.002	.002
Bank by Select	214.77	.007	.007
Bank by Stop	1030.04	.034	.034
Select by Stop	84.24	.003	.003
Cor by Bank by Select	10.82	.000	.000
Cor by Bank by Stop	1.21	.000	.000
Cor by Select by Stop	4.52	.000	.000
Bank by Select by Stop	40.64	.001	.001
Residuals	16.31		
Total	30011.36		

Table B.25: The sums of squares (Sum Sq.), $\eta^2 = \frac{\text{SSF}}{\text{SST}}$, and $\omega^2 = \frac{\text{SSF} - \text{df}_F \times \text{MSE}}{\text{SST} + \text{MSE}}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting average loss (with $P = 100$) given a compensatory classification bound function. The ANOVA was run with all main effects, two-way interactions, and three-way interactions.

Variance Type	Sum Sq.	η^2	ω^2
Correlation (Cor)	1596.31	.055	.055
Item Bank (Bank)	8713.36	.300	.300
Select Alg. (Select)	458.52	.016	.016
Stop Rule (Stop)	16867.19	.581	.581
Cor by Bank	23.51	.001	.001
Cor by Select	8.50	.000	.000
Cor by Stop	40.56	.001	.001
Bank by Select	274.68	.009	.009
Bank by Stop	910.04	.031	.031
Select by Stop	27.23	.003	.002
Cor by Bank by Select	12.40	.000	.000
Cor by Bank by Stop	3.21	.000	.000
Cor by Select by Stop	8.04	.000	.000
Bank by Select by Stop	30.39	.001	.001
Residuals	30.68		
Total	29050.62		

Table B.26: The sums of squares (Sum Sq.), $\eta^2 = \frac{\text{SSF}}{\text{SST}}$, and $\omega^2 = \frac{\text{SSF} - \text{df}_F \times \text{MSE}}{\text{SST} + \text{MSE}}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting average loss (with $P = 500$) given a compensatory classification bound function. The ANOVA was run with all main effects, two-way interactions, and three-way interactions.

Variance Type	Sum Sq.	η^2	ω^2
Correlation (Cor)	5873.68	.171	.171
Item Bank (Bank)	18071.32	.527	.527
Select Alg. (Select)	871.64	.025	.025
Stop Rule (Stop)	7357.64	.215	.214
Cor by Bank	37.78	.001	.001
Cor by Select	26.00	.001	.000
Cor by Stop	24.33	.001	.000
Bank by Select	602.59	.018	.017
Bank by Stop	652.44	.019	.018
Select by Stop	162.14	.005	.002
Cor by Bank by Select	31.25	.001	.001
Cor by Bank by Stop	39.93	.001	.001
Cor by Select by Stop	97.63	.003	.000
Bank by Select by Stop	95.21	.003	.000
Residuals	333.79		
Total	34277.35		

Table B.27: The sums of squares (Sum Sq.), $\eta^2 = \frac{\text{SSF}}{\text{SST}}$, and $\omega^2 = \frac{\text{SSF} - \text{df}_F \times \text{MSE}}{\text{SST} + \text{MSE}}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting average loss (with $P = 1000$) given a compensatory classification bound function. The ANOVA was run with all main effects, two-way interactions, and three-way interactions.

Variance Type	Sum Sq.	η^2	ω^2
Correlation (Cor)	15005.63	.245	.245
Item Bank (Bank)	34516.04	.564	.564
Select Alg. (Select)	1629.49	.027	.026
Stop Rule (Stop)	4986.82	.081	.080
Cor by Bank	60.35	.001	.001
Cor by Select	65.33	.001	.000
Cor by Stop	47.01	.001	.000
Bank by Select	1211.06	.020	.019
Bank by Stop	830.79	.014	.012
Select by Stop	572.43	.009	.004
Cor by Bank by Select	82.96	.001	.001
Cor by Bank by Stop	150.41	.002	.001
Cor by Select by Stop	379.57	.006	.001
Bank by Select by Stop	414.33	.007	.001
Residuals	1265.37		
Total	61217.58		

Table B.28: The sums of squares (Sum Sq.), $\eta^2 = \frac{\text{SSF}}{\text{SST}}$, and $\omega^2 = \frac{\text{SSF} - \text{df}_F \times \text{MSE}}{\text{SST} + \text{MSE}}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting mean classification accuracy given a non-compensatory classification bound function. The ANOVA was run with all main effects, two-way interactions, and three-way interactions.

Variance Type	Sum Sq.	η^2	ω^2
Correlation (Cor)	.00228	.014	.014
Item Bank (Bank)	.00202	.013	.012
Select Alg. (Select)	.01002	.062	.062
Stop Rule (Stop)	.04586	.285	.283
Cor by Bank	.00033	.002	.002
Cor by Select	.00052	.003	.003
Cor by Stop	.00151	.009	.008
Bank by Select	.00922	.057	.057
Bank by Stop	.01417	.088	.087
Select by Stop	.03695	.229	.224
Cor by Bank by Select	.00008	.000	.000
Cor by Bank by Stop	.00227	.014	.013
Cor by Select by Stop	.00207	.013	.007
Bank by Select by Stop	.03036	.188	.183
Residuals	.00343		
Total	.16110		

Table B.29: The sums of squares (Sum Sq.), $\eta^2 = \frac{\text{SSF}}{\text{SST}}$, and $\omega^2 = \frac{\text{SSF} - \text{df}_F \times \text{MSE}}{\text{SST} + \text{MSE}}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting mean test length given a non-compensatory classification bound function. The ANOVA was run with all main effects, two-way interactions, and three-way interactions.

Variance Type	Sum Sq.	η^2	ω^2
Correlation (Cor)	23.52	.000	.000
Item Bank (Bank)	9372.46	.198	.198
Select Alg. (Select)	2499.74	.053	.053
Stop Rule (Stop)	23168.56	.488	.488
Cor by Bank	31.34	.001	.001
Cor by Select	11.53	.000	.000
Cor by Stop	44.09	.001	.001
Bank by Select	529.32	.011	.011
Bank by Stop	8308.66	.175	.175
Select by Stop	2147.33	.045	.045
Cor by Bank by Select	3.43	.000	.000
Cor by Bank by Stop	20.03	.000	.000
Cor by Select by Stop	45.52	.001	.001
Bank by Select by Stop	1166.62	.025	.024
Residuals	63.67		
Total	47435.8		

Table B.30: The sums of squares (Sum Sq.,), $\eta^2 = \frac{\text{SSF}}{\text{SST}}$, and $\omega^2 = \frac{\text{SSF} - \text{df}_F \times \text{MSE}}{\text{SST} + \text{MSE}}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting average loss (with $P = 100$) given a non-compensatory classification bound function. The ANOVA was run with all main effects, two-way interactions, and three-way interactions.

Variance Type	Sum Sq.	η^2	ω^2
Correlation (Cor)	92.71	.002	.002
Item Bank (Bank)	10261.88	.232	.232
Select Alg. (Select)	1850.14	.042	.042
Stop Rule (Stop)	20898.56	.473	.473
Cor by Bank	14.39	.000	.000
Cor by Select	12.80	.000	.000
Cor by Stop	58.93	.001	.001
Bank by Select	655.37	.015	.015
Bank by Stop	7826.82	.177	.177
Select by Stop	1588.97	.036	.035
Cor by Bank by Select	3.43	.000	.000
Cor by Bank by Stop	48.70	.001	.001
Cor by Select by Stop	59.37	.001	.001
Bank by Select by Stop	699.06	.016	.015
Residuals	86.11		
Total	44157.22		

Table B.31: The sums of squares (Sum Sq.), $\eta^2 = \frac{\text{SSF}}{\text{SST}}$, and $\omega^2 = \frac{\text{SSF} - \text{df}_F \times \text{MSE}}{\text{SST} + \text{MSE}}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting average loss (with $P = 500$) given a non-compensatory classification bound function. The ANOVA was run with all main effects, two-way interactions, and three-way interactions.

Variance Type	Sum Sq.	η^2	ω^2
Correlation (Cor)	826.27	.013	.013
Item Bank (Bank)	14222.67	.225	.225
Select Alg. (Select)	1255.69	.020	.019
Stop Rule (Stop)	20990.53	.332	.331
Cor by Bank	11.74	.000	.000
Cor by Select	122.77	.002	.002
Cor by Stop	420.67	.007	.006
Bank by Select	3004.08	.047	.047
Bank by Stop	8733.25	.138	.137
Select by Stop	6746.22	.107	.103
Cor by Bank by Select	19.20	.000	.000
Cor by Bank by Stop	616.90	.010	.009
Cor by Select by Stop	528.21	.008	.005
Bank by Select by Stop	4901.73	.077	.074
Residuals	862.58		
Total	63262.51		

Table B.32: The sums of squares (Sum Sq.), $\eta^2 = \frac{\text{SSF}}{\text{SST}}$, and $\omega^2 = \frac{\text{SSF} - \text{df}_F \times \text{MSE}}{\text{SST} + \text{MSE}}$, where SSF is the sums of squares for a particular factor and df_F is the corresponding degrees of freedom, for an ANOVA predicting average loss (with $P = 1000$) given a non-compensatory classification bound function. The ANOVA was run with all main effects, two-way interactions, and three-way interactions.

Variance Type	Sum Sq.	η^2	ω^2
Correlation (Cor)	2771.02	.017	.017
Item Bank (Bank)	20080.65	.126	.126
Select Alg. (Select)	5021.52	.031	.031
Stop Rule (Stop)	41742.42	.261	.260
Cor by Bank	155.06	.001	.001
Cor by Select	496.26	.003	.002
Cor by Stop	1553.18	.010	.008
Bank by Select	10090.19	.063	.063
Bank by Stop	16242.32	.102	.100
Select by Stop	29821.84	.187	.181
Cor by Bank by Select	74.37	.000	.000
Cor by Bank by Stop	2347.48	.015	.013
Cor by Select by Stop	2044.56	.013	.007
Bank by Select by Stop	23819.22	.149	.144
Residuals	3378.25		
Total	159638.2		

Appendix C

Figures: Aggregate over Distribution

The following figures depict the relationship between accuracy and test length for various conditions aggregated across a distribution of simulees. In all of the figures, ability was simulated from a bivariate normal distribution.

C.1 Scatterplots

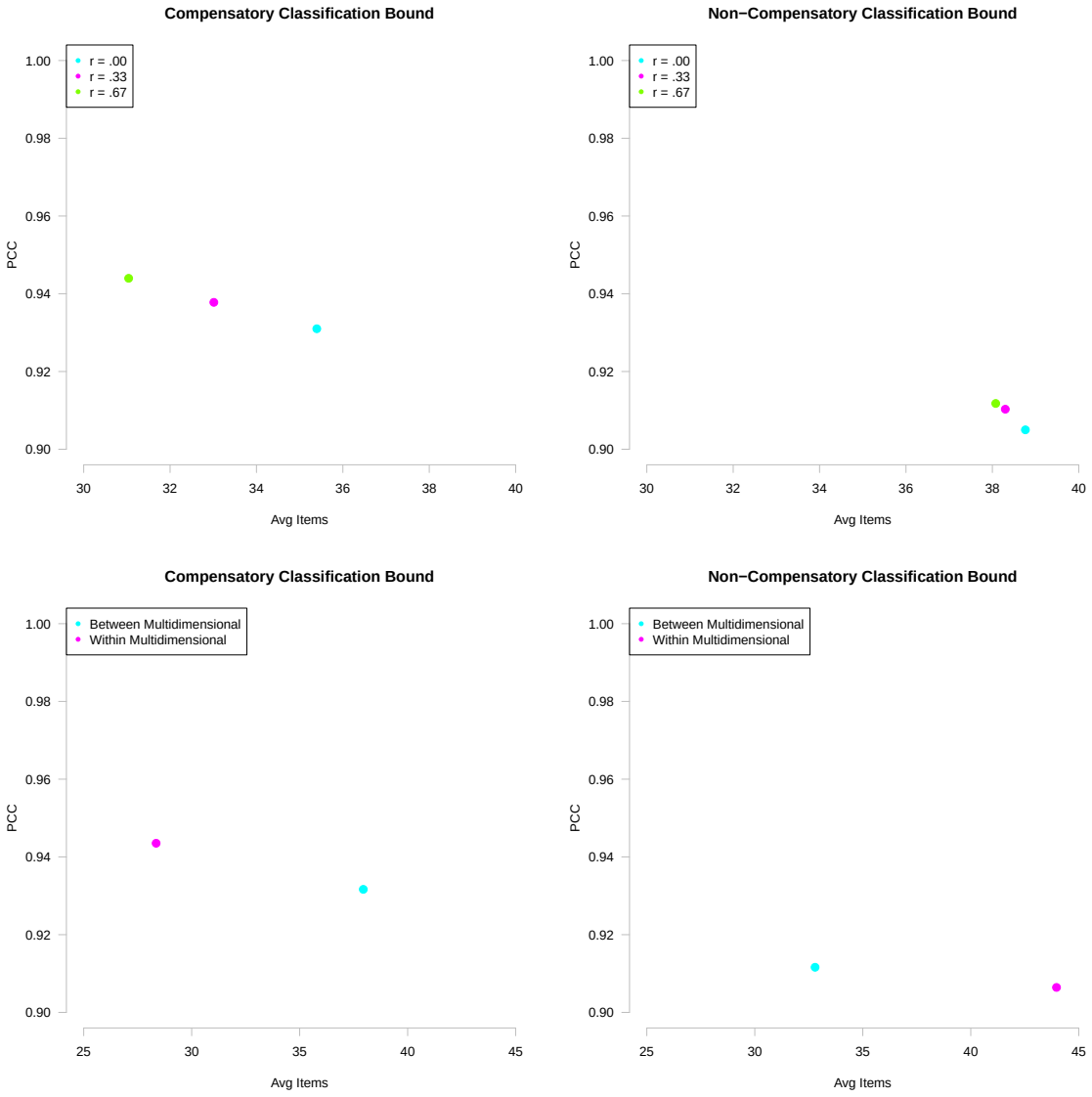

Figure C.1: Scatterplots of the percent classified correctly (PCC) by average number of items administered for different true correlations between ability dimensions (top panel) and different item banks (bottom panel) using either a compensatory classification bound function (left panel) or a non-compensatory classification bound function (right panels).

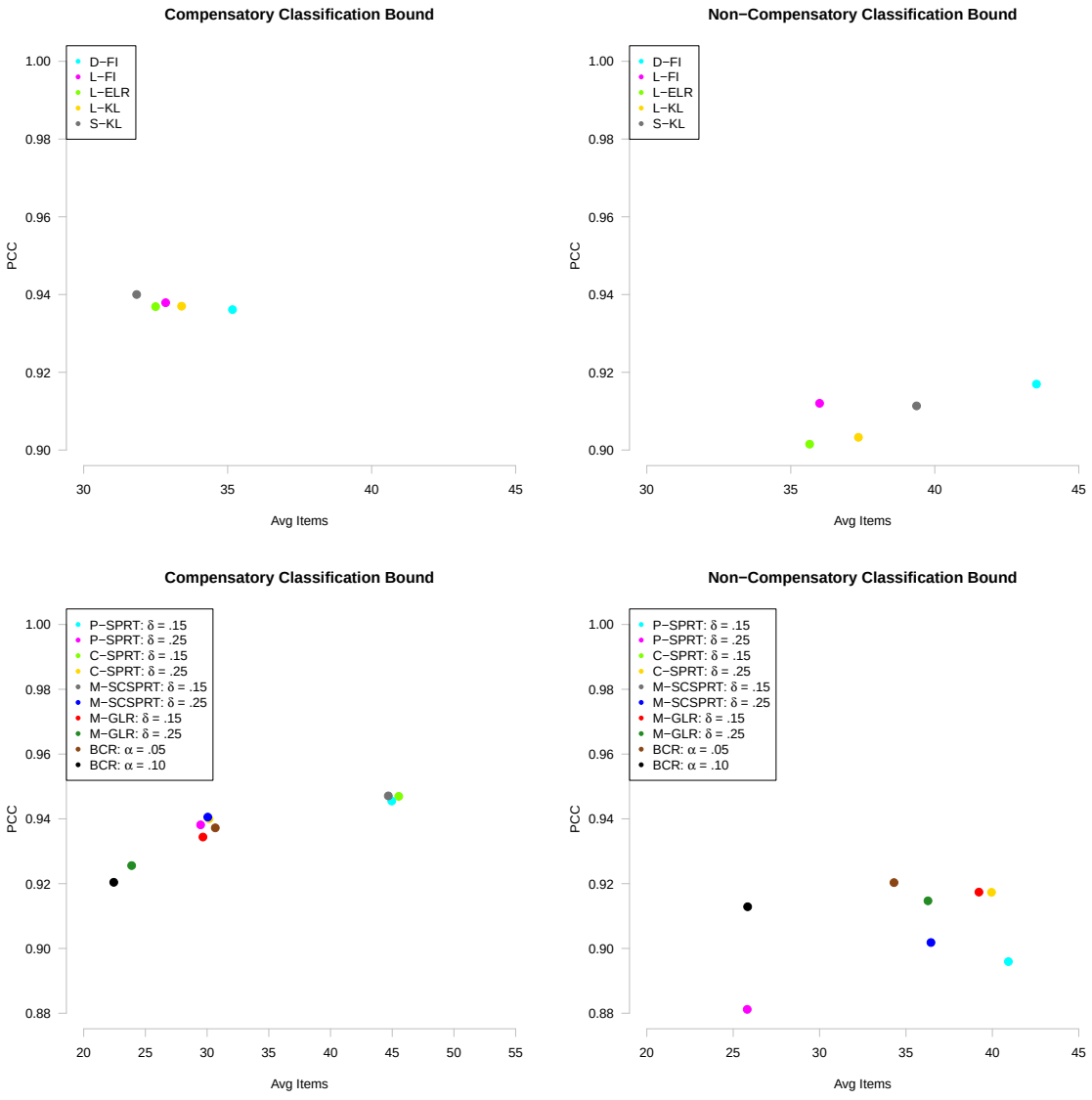

Figure C.2: Scatterplots of the percent classified correctly (PCC) by average number of items administered for different item selection algorithms (top panel) and different stopping rules (bottom panels) using either a compensatory classification bound function (left panel) or a non-compensatory classification bound function (right panels).

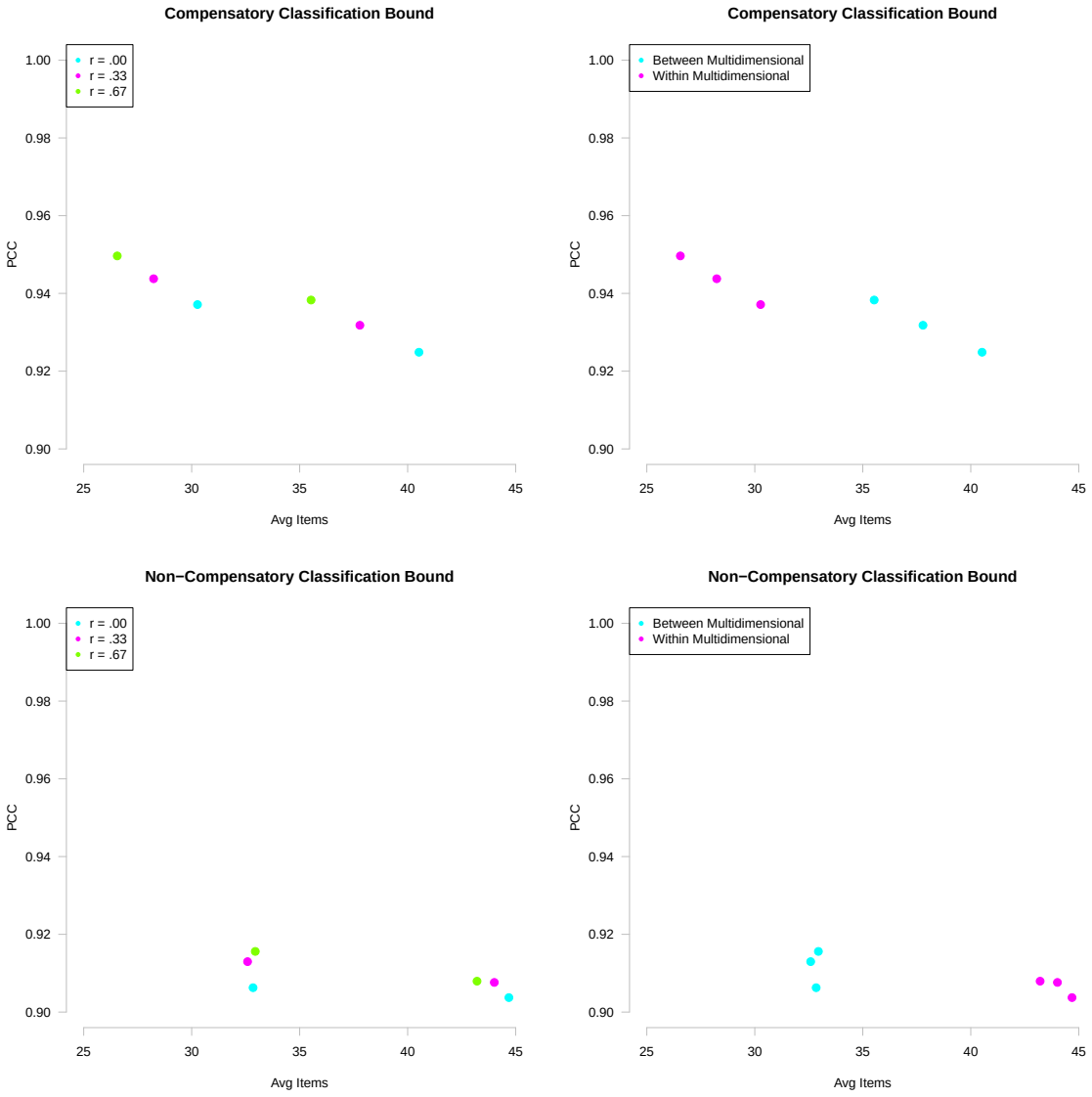

Figure C.3: Scatterplots of the percent classified correctly (PCC) by average number of items administered based on the interaction between true correlations between ability dimensions and item bank using either a compensatory classification bound function (top panels) or a non-compensatory classification bound function (bottom panels). The left panels are color coded according to correlation condition, whereas the right panels are color coded according to item bank.

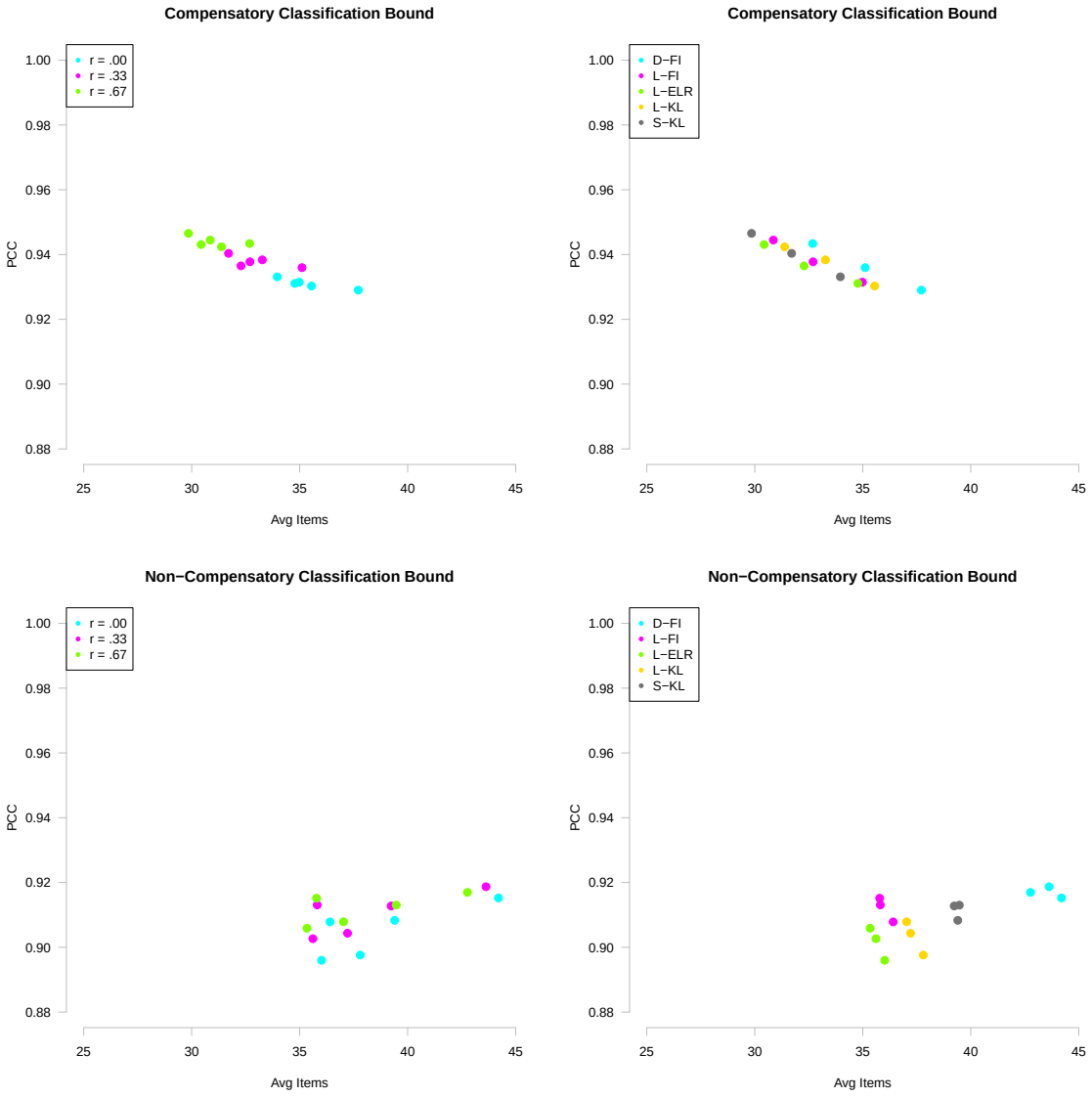

Figure C.4: Scatterplots of the percent classified correctly (PCC) by average number of items administered based on the interaction between true correlations between ability dimensions and item selection algorithm using either a compensatory classification bound function (top panels) or a non-compensatory classification bound function (bottom panels). The left panels are color coded according to correlation condition, whereas the right panels are color coded according to item selection algorithm.

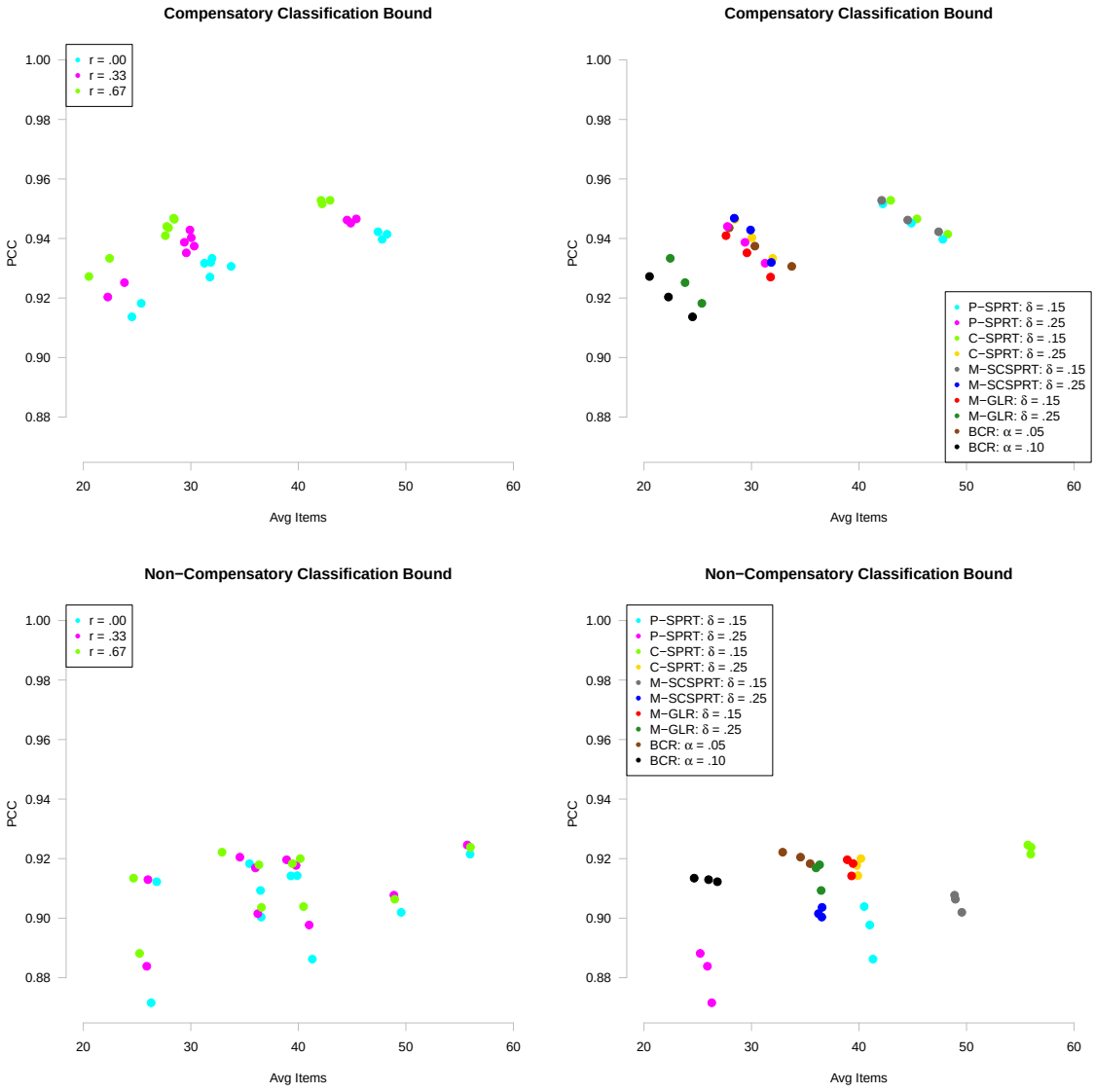

Figure C.5: Scatterplots of the percent classified correctly (PCC) by average number of items administered based on the interaction between true correlations between ability dimensions and stopping rule using either a compensatory classification bound function (top panels) or a non-compensatory classification bound function (bottom panels). The left panels are color coded according to correlation condition, whereas the right panels are color coded according to stopping rule.

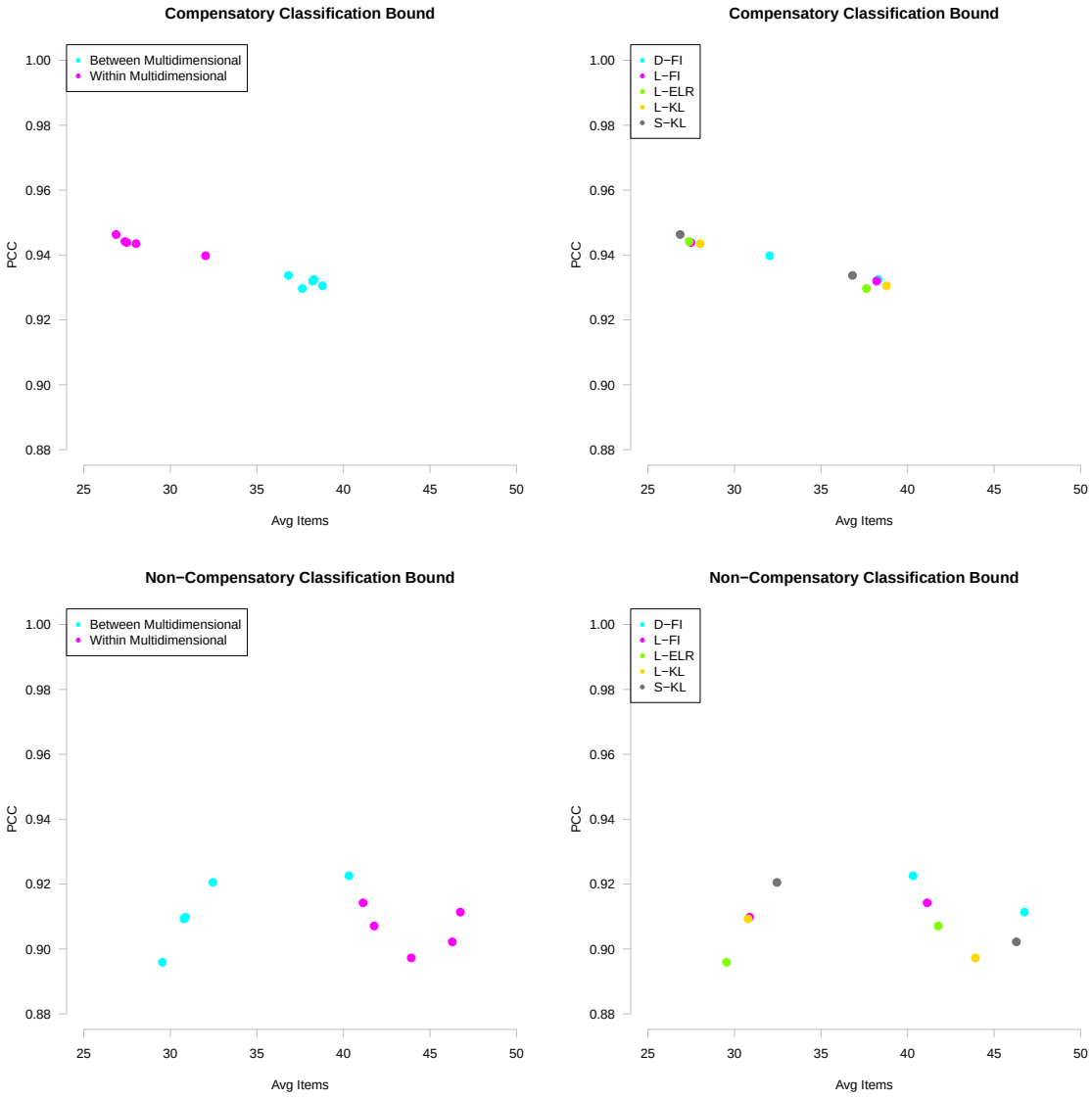

Figure C.6: Scatterplots of the percent classified correctly (PCC) by average number of items administered based on the interaction between item bank and item selection algorithm using either a compensatory classification bound function (top panels) or a non-compensatory classification bound function (bottom panels). The left panels are color coded according to item bank, whereas the right panels are color coded according to item selection algorithm.

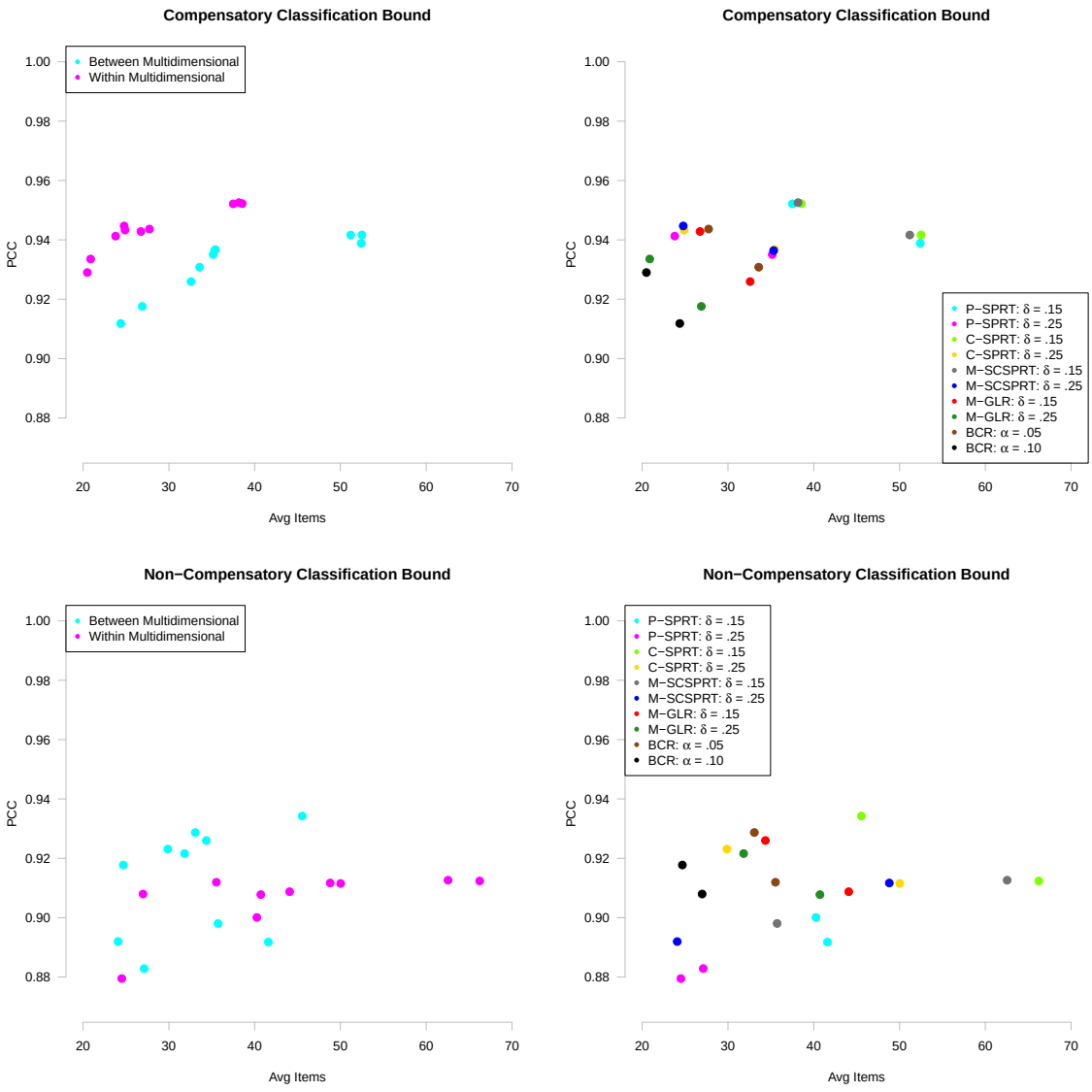

Figure C.7: Scatterplots of the percent classified correctly (PCC) by average number of items administered based on the interaction between item bank and stopping rule using either a compensatory classification bound function (top panels) or a non-compensatory classification bound function (bottom panels). The left panels are color coded according to item bank, whereas the right panels are color coded according to stopping rule.

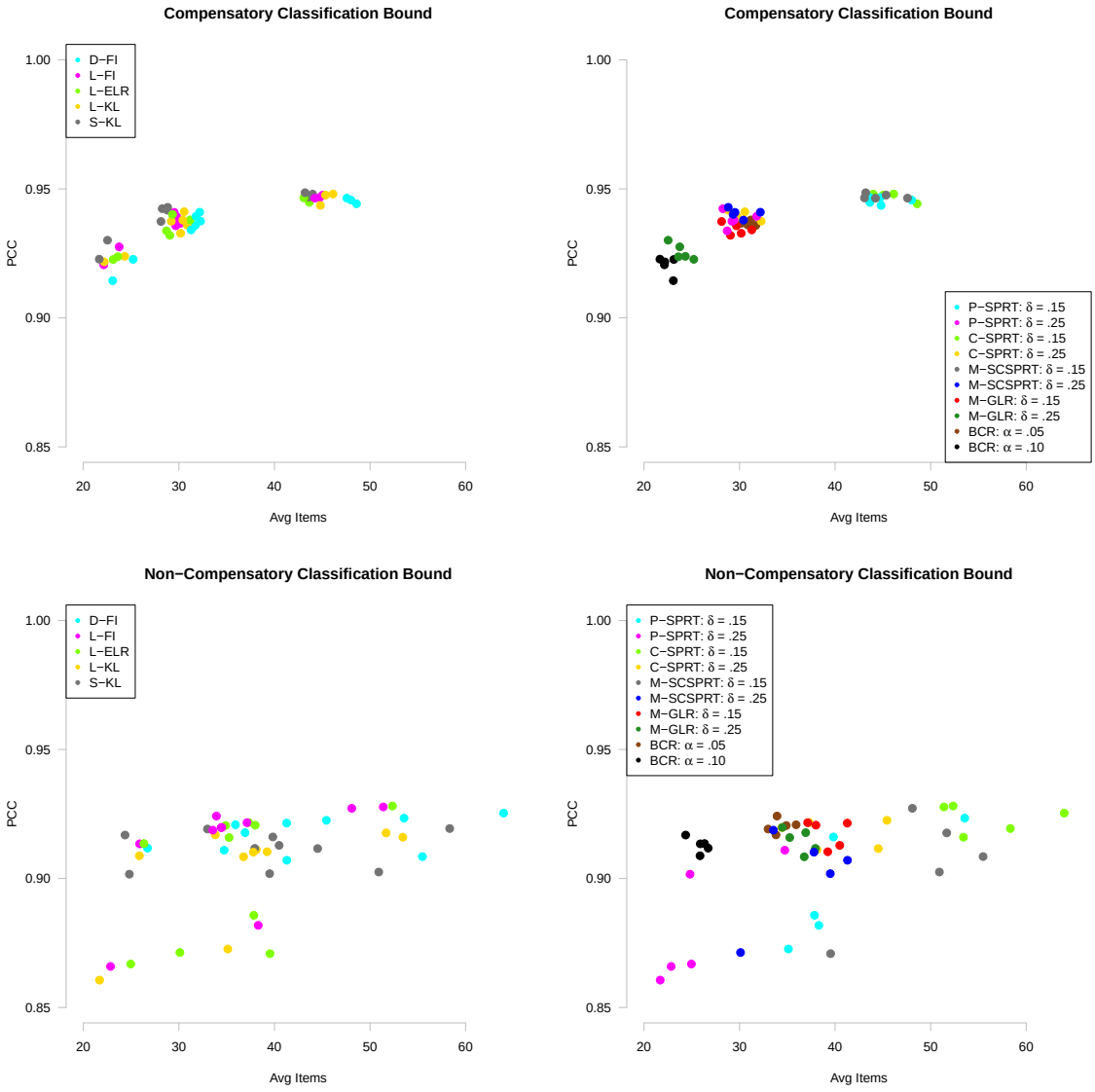

Figure C.8: Scatterplots of the percent classified correctly (PCC) by average number of items administered based on the interaction between item selection algorithm and stopping rule using either a compensatory classification bound function (top panels) or a non-compensatory classification bound function (bottom panels). The left panels are color coded according to item selection algorithm, whereas the right panels are color coded according to stopping rule.

C.2 Loss Trend Plots

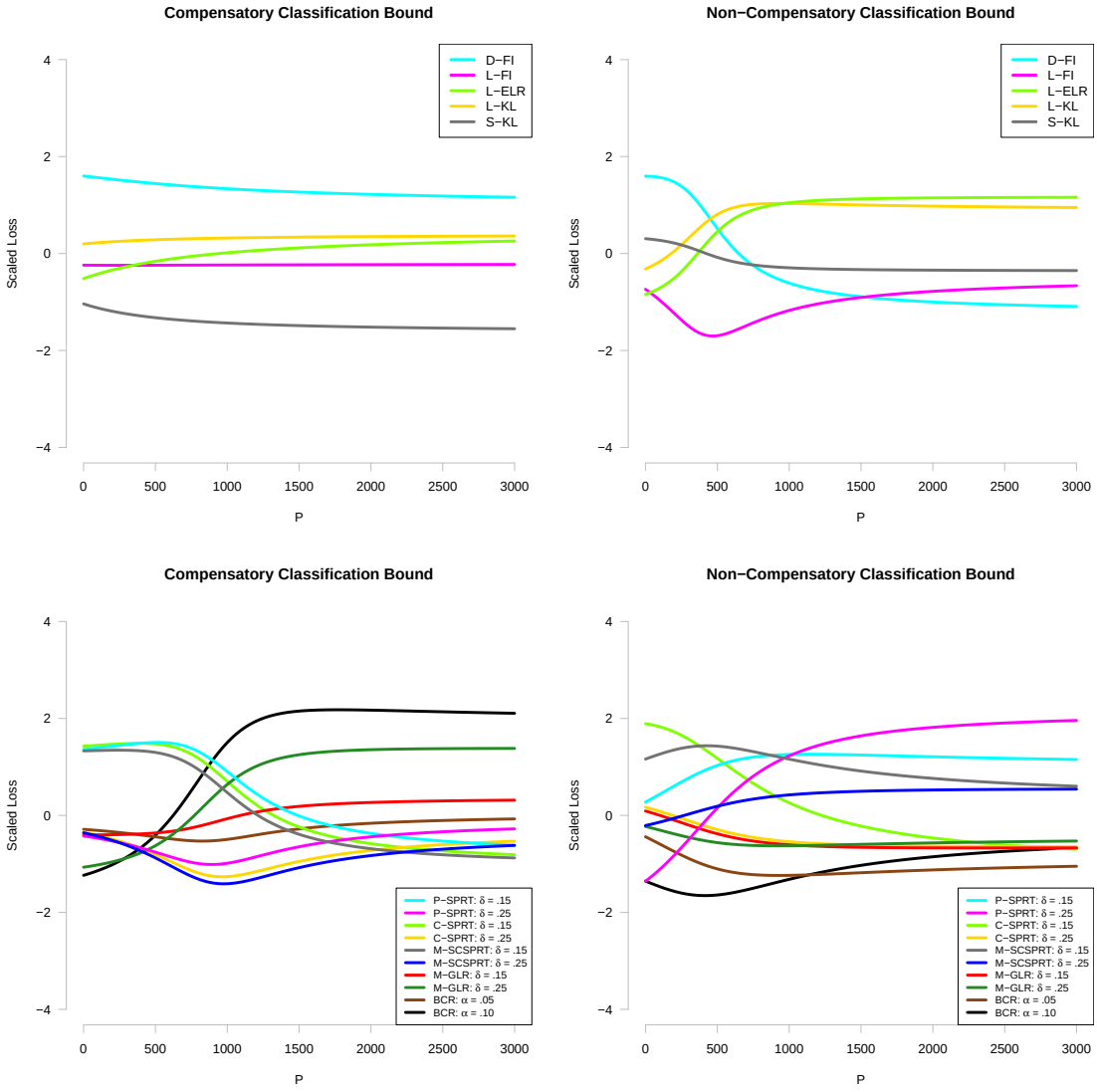

Figure C.9: Average loss within each item selection algorithm or stopping rule for various values of P, where Loss $= P \times I_W + J$ (see Appendix B). The upper panels indicate the average loss for each of the item selection algorithms, whereas the lower panels indicate the average loss for each of the stopping rules. The left panels represent a compensatory classification bound function, whereas the right panels represent a non-compensatory classification bound function.

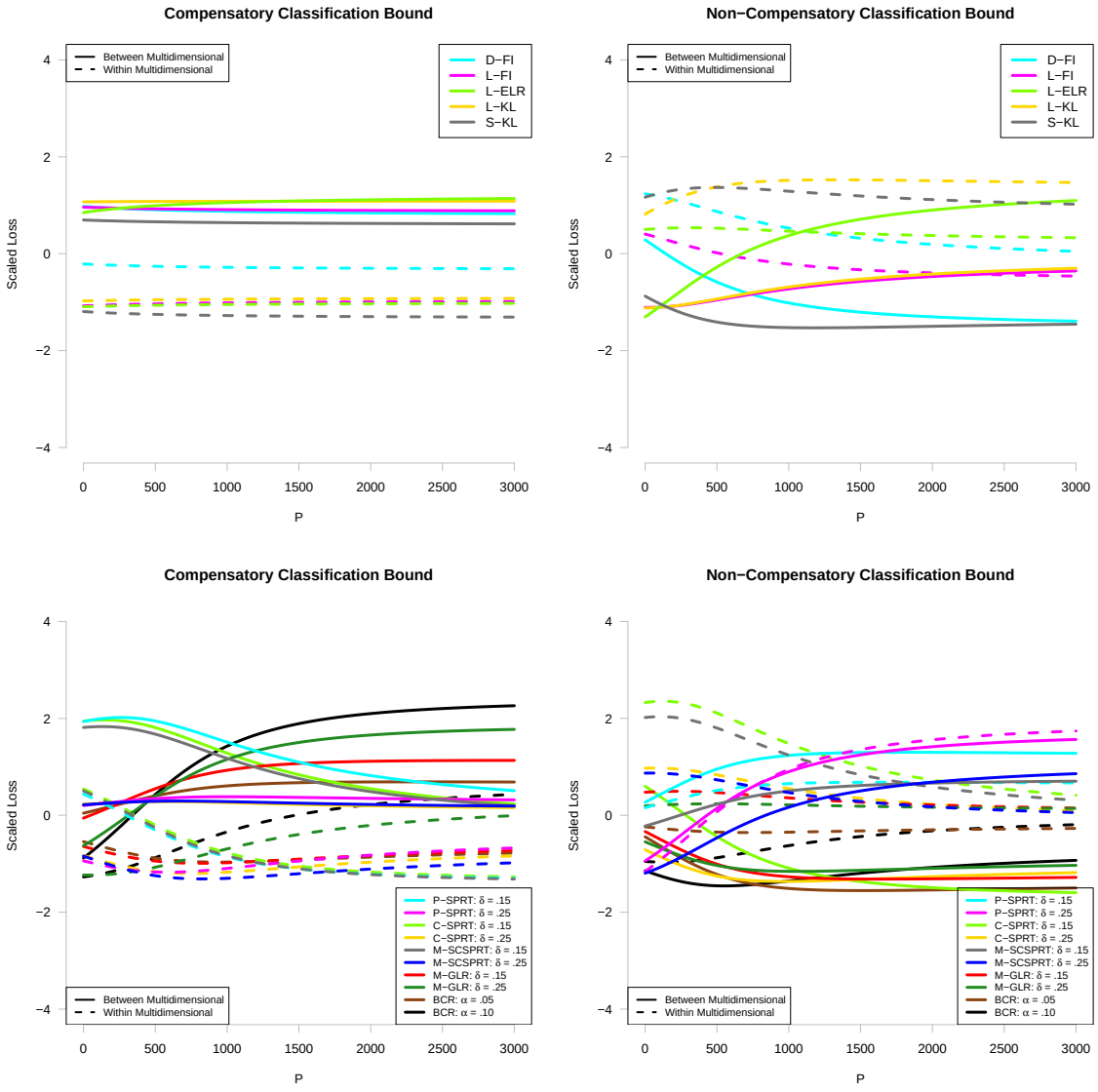

Figure C.10: Average loss within each item selection algorithm by item bank or stopping rule by item bank for various values of P, where Loss $= P \times I_W + J$ (see Appendix B). The upper panels indicate the average loss for each of the item selection algorithms by item bank, whereas the lower panels indicate the average loss for each of the stopping rules by item bank. The left panels represent a compensatory classification bound function, whereas the right panels represent a non-compensatory classification bound function. Colors are coded according to item selection algorithm or stopping rule, whereas line type is determined by item bank.

Appendix D

Figures: Conditional on Ability

The following figures depict the accuracy, test length, and loss conditional on various ability vectors for 48 out of the 600 overall conditions. These conditions were chosen by examining the scatterplots and loss trend plots when aggregating over a distribution. I simplified the number of conditions as follows: (1) The true correlation between θ_1 and θ_2 was always assumed to be .33; (2) When using the compensatory classification bound function, the MCMT algorithm selected items from the within-item multidimensional bank; (3) When using the non-compensatory classification bound function, the MCMT algorithm selected items from the between-item multidimensional bank; and (4) The weakest item selection algorithm (D-FI) and stopping rule (P-SPRT) were eliminated. Note that loss is defined as the average of Loss $= P \times I_W + J$, where I_W is an indicator function for incorrect classification, J is the number of items given to an examinee, and P is the penalty accrued for an incorrect decision. In all cases, I chose $P = 500$.

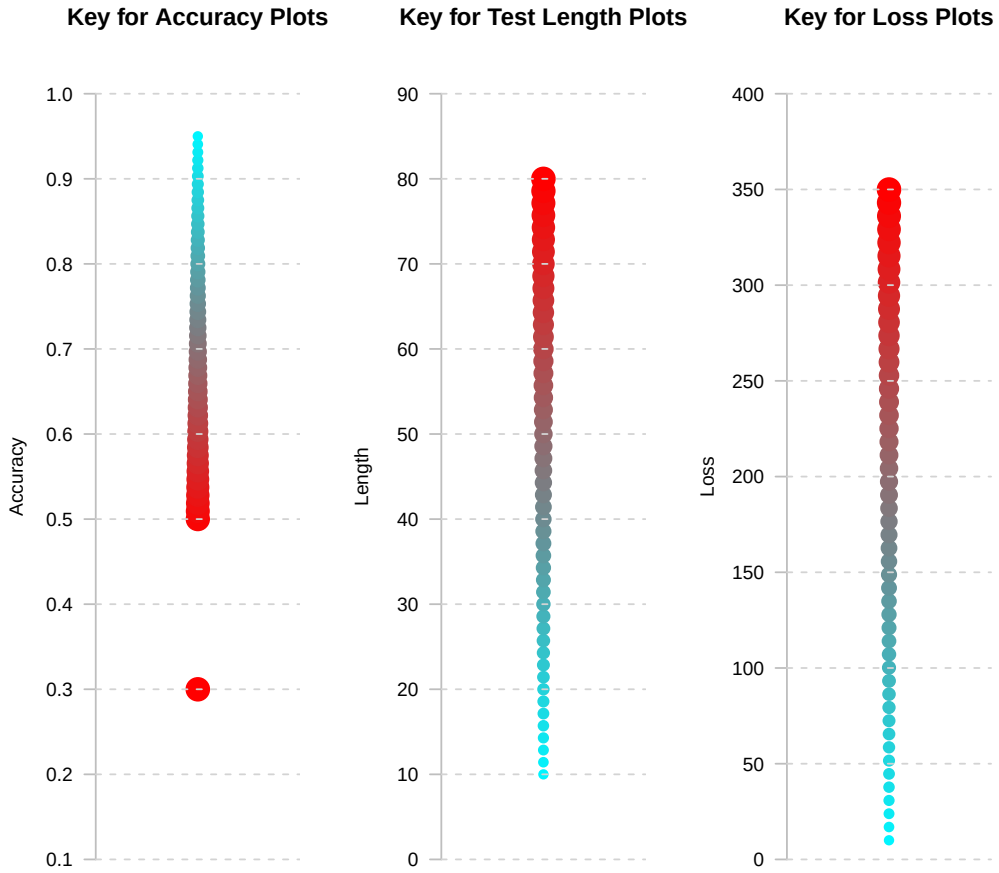

Figure D.1: Legends for the conditional accuracy, test length, and loss function plots depicted on the following pages. The left-most panel indicates the bubble color and size code for the accuracy plots, the middle panel depicts the same information for the test length plots, and the right-most panel depicts the same information for the loss function plots. In all cases, small blue points represent good results, whereas large red points represent bad results.

D.1 Accuracy Plots

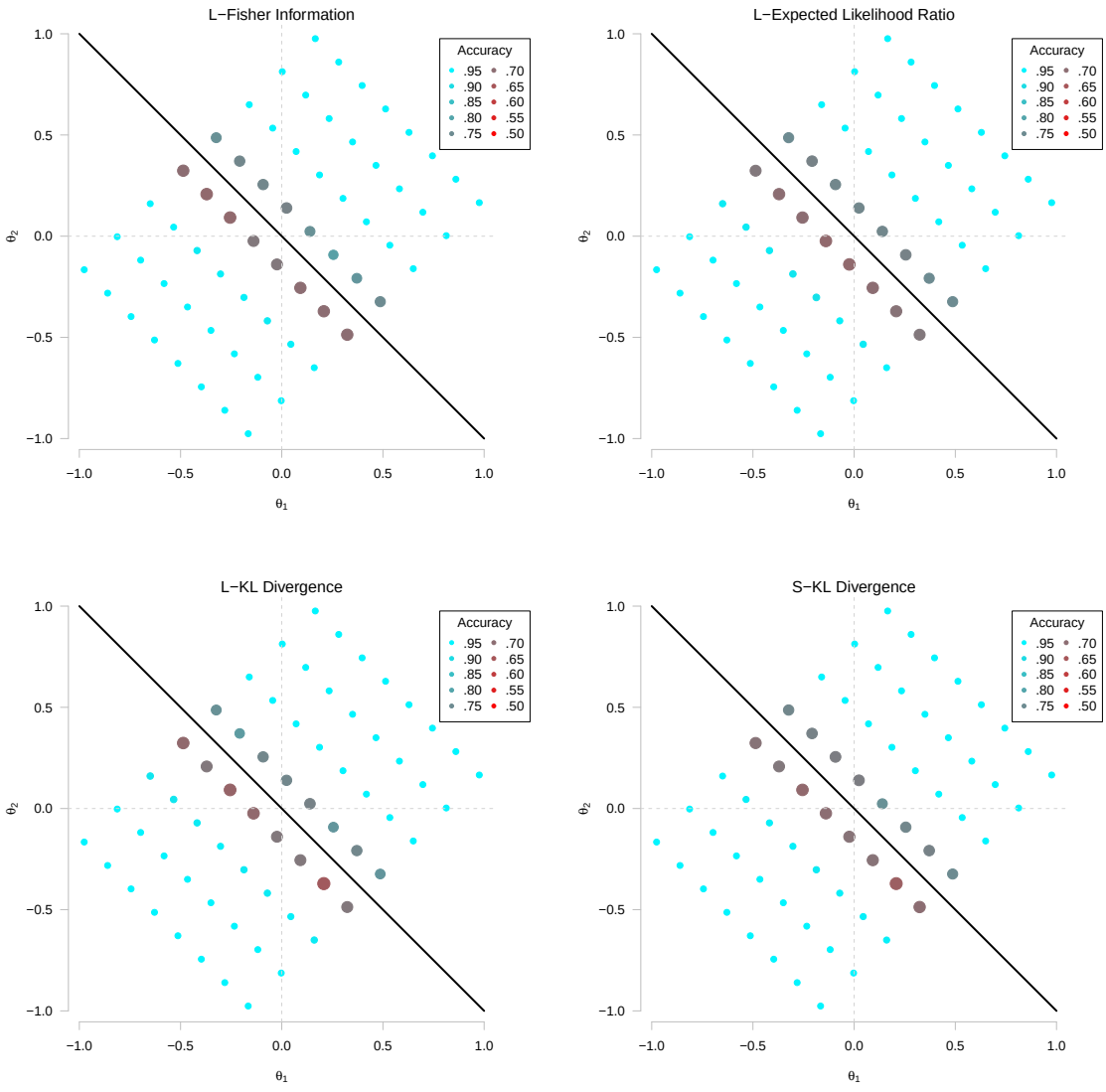

Figure D.2: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the compensatory classification bound function and the C-SPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

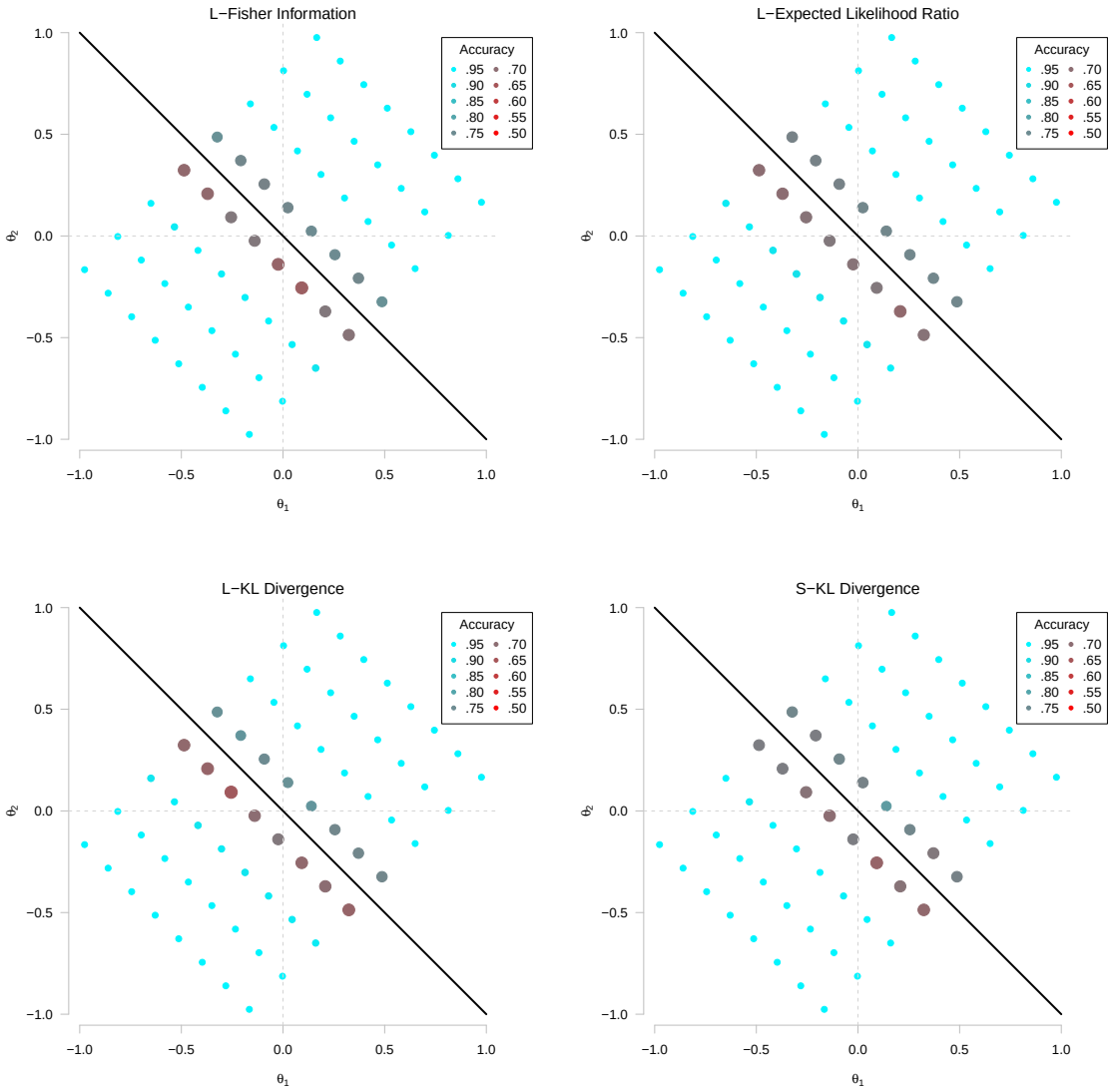

Figure D.3: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the compensatory classification bound function and the M-SCSPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

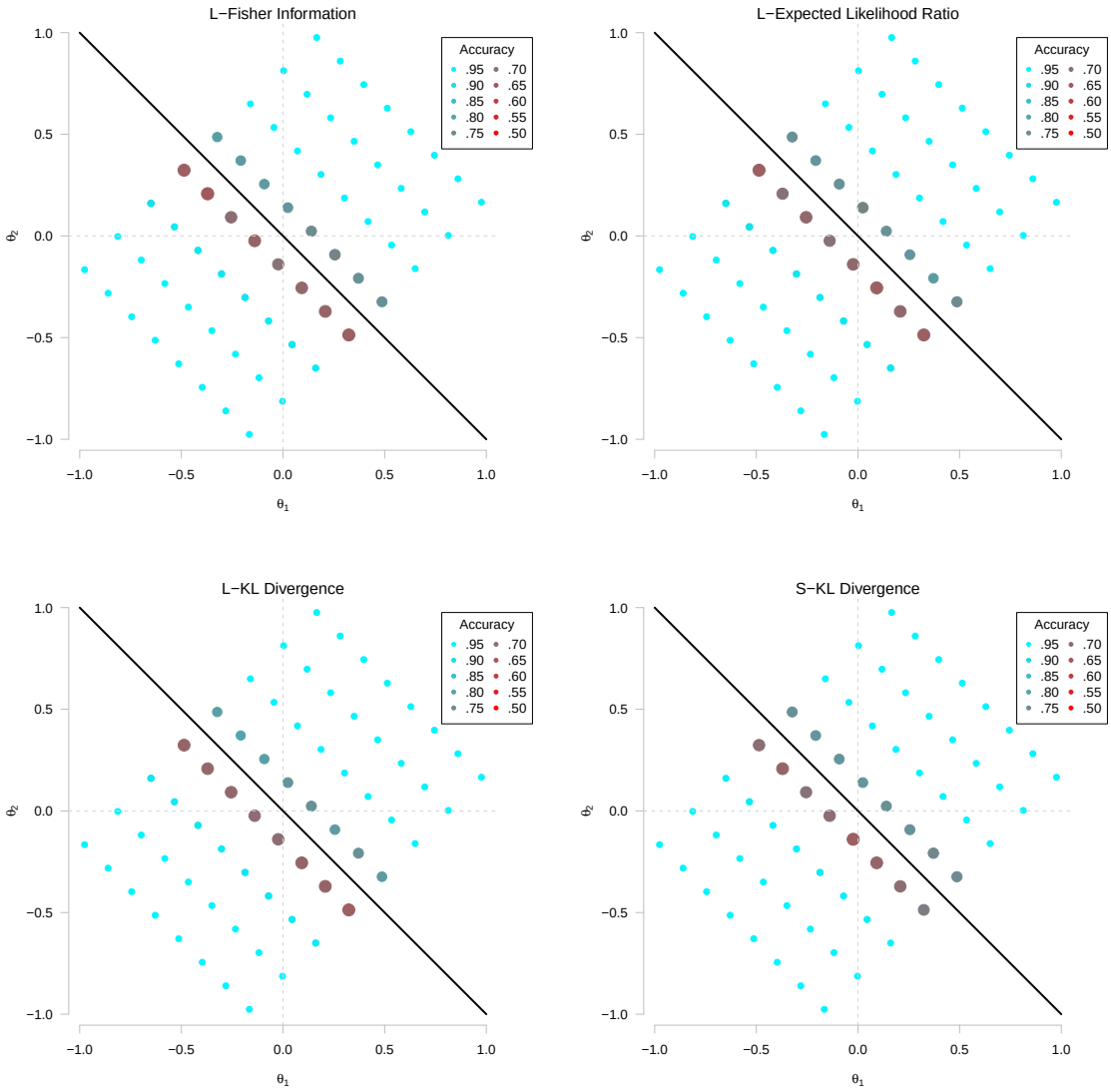

Figure D.4: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the compensatory classification bound function and the M-GLR stopping rule with $\delta = .15$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

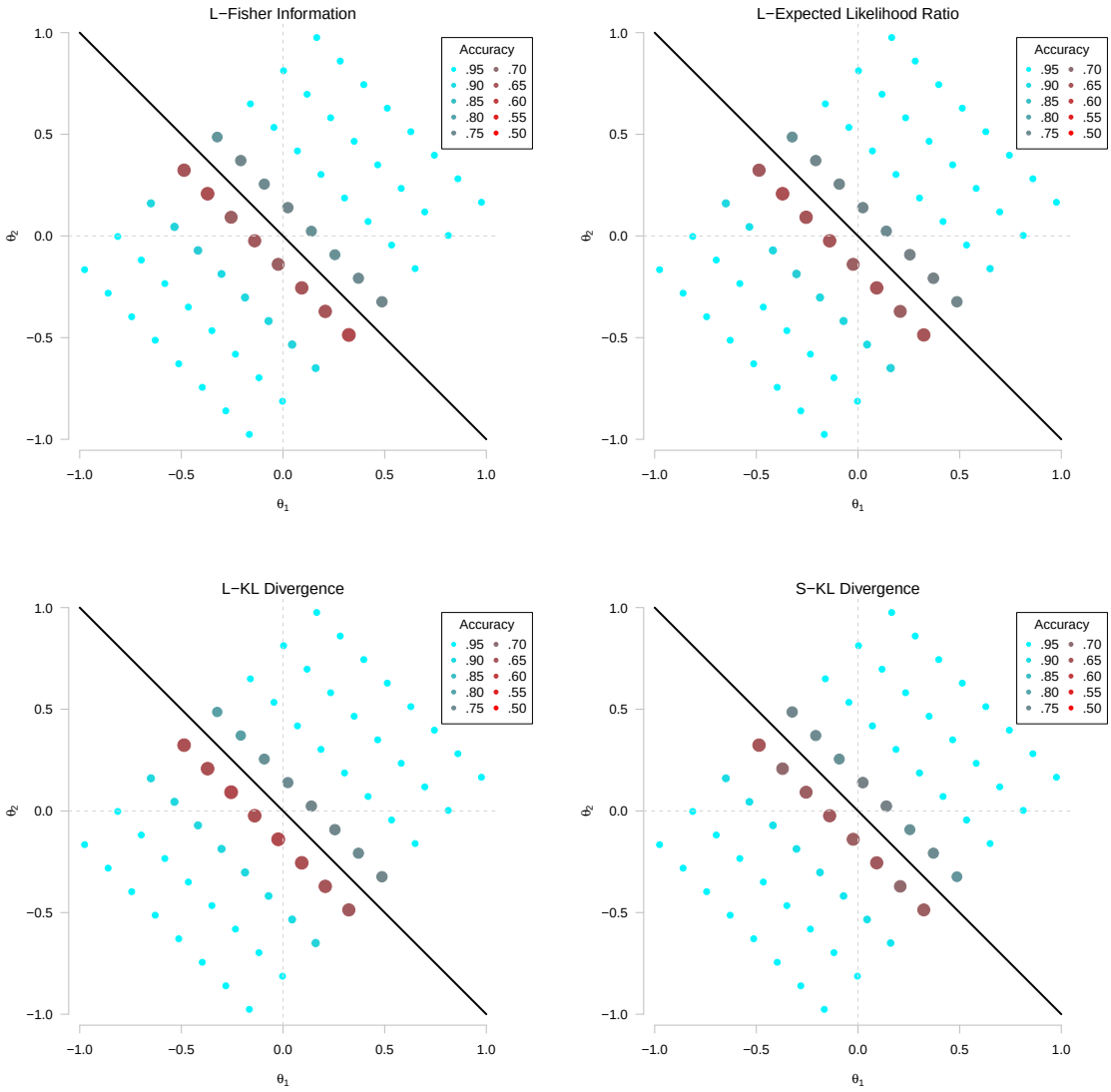

Figure D.5: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the compensatory classification bound function and the M-GLR stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

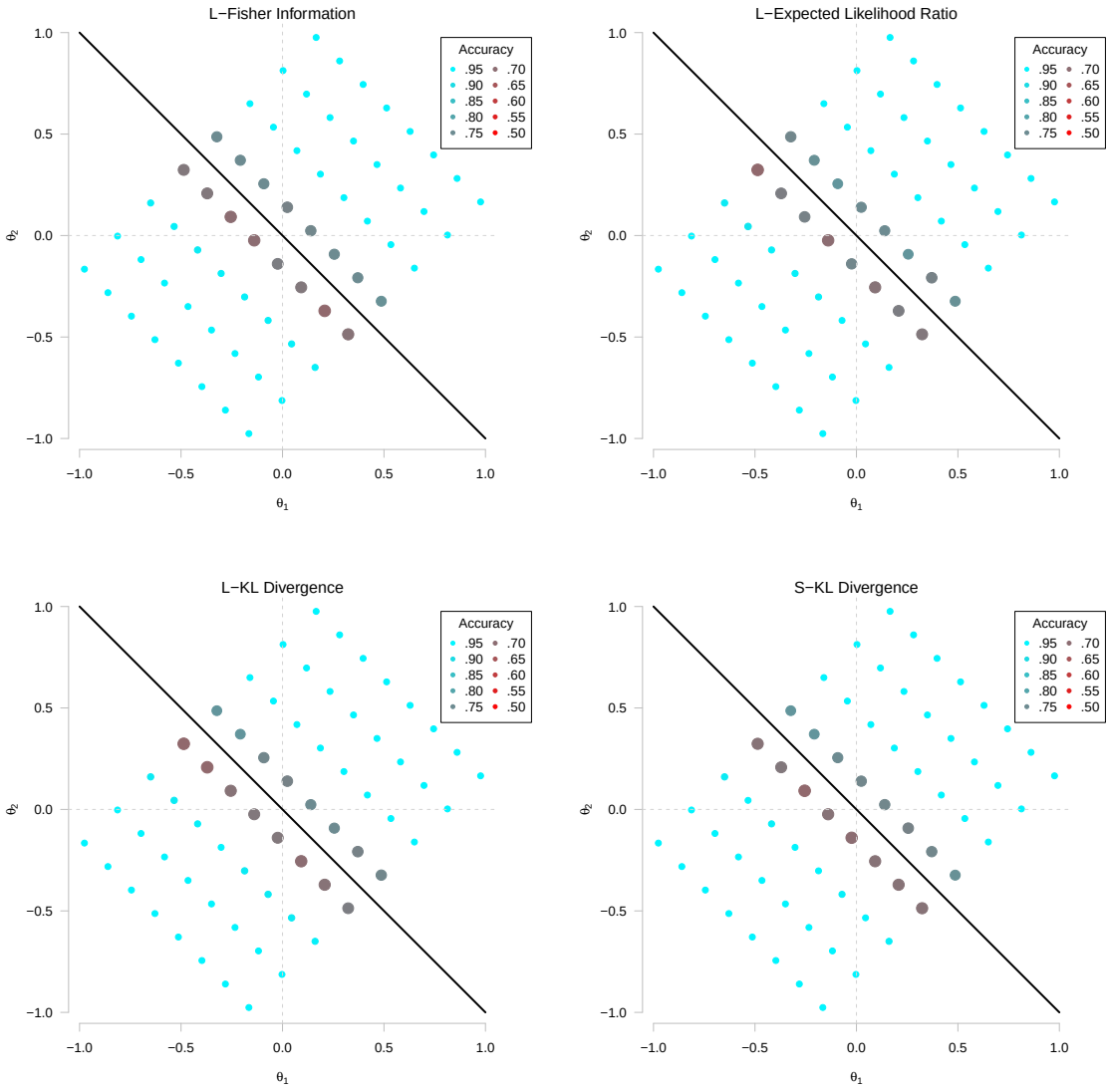

Figure D.6: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the compensatory classification bound function and the BCR stopping rule with $\alpha = .05$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

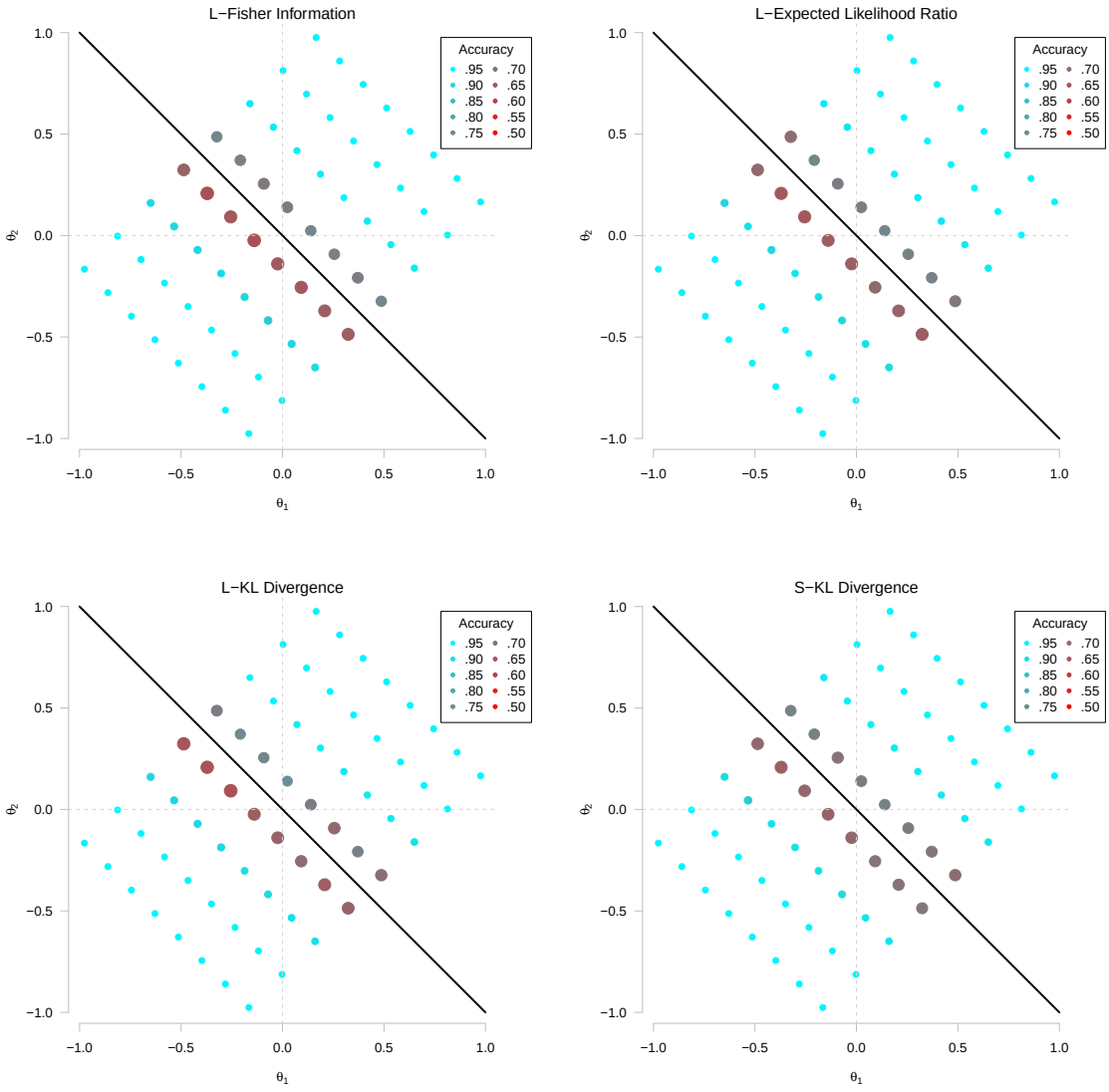

Figure D.7: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the compensatory classification bound function and the BCR stopping rule with $\alpha = .10$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

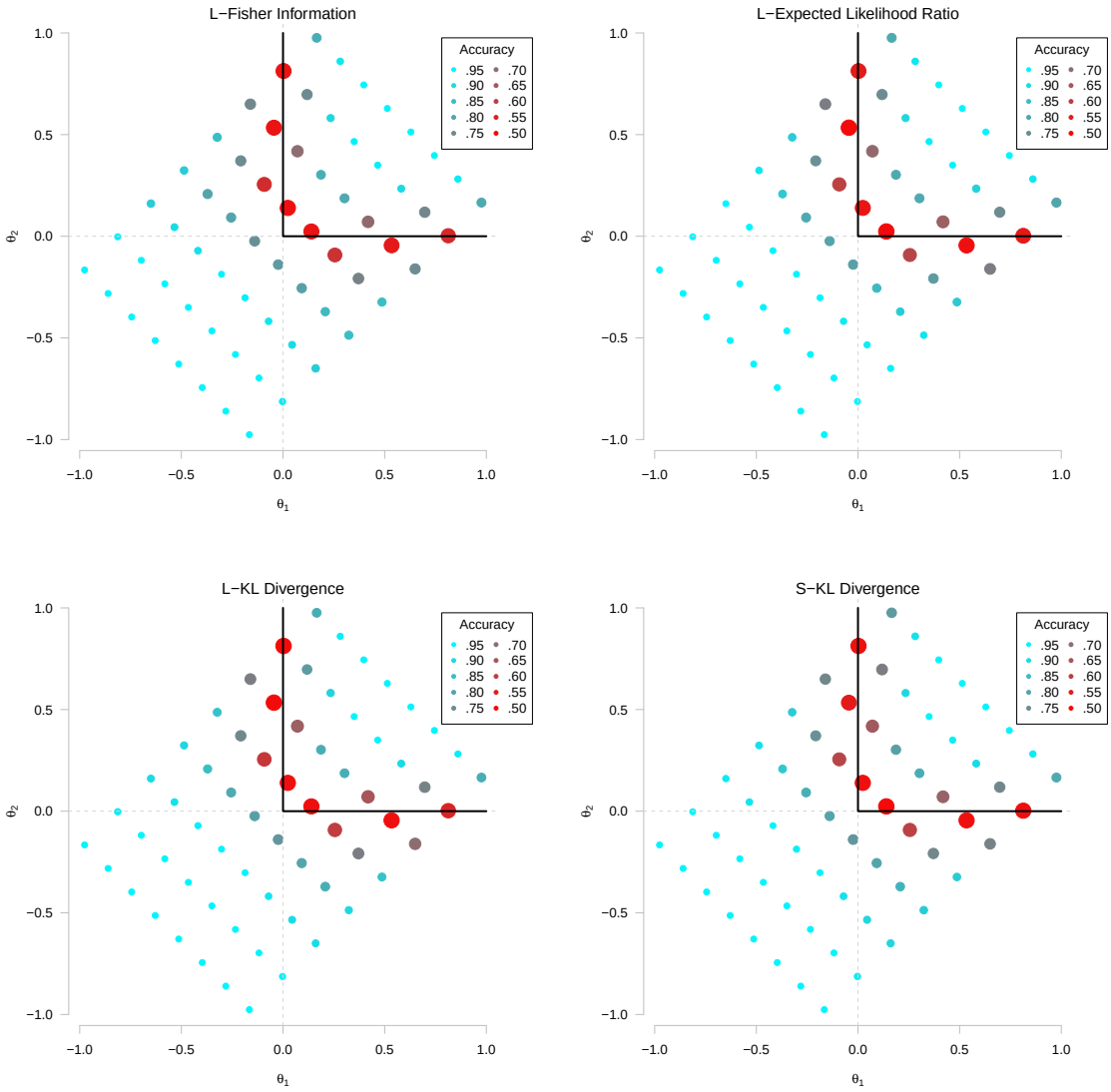

Figure D.8: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the non-compensatory classification bound function and the C-SPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

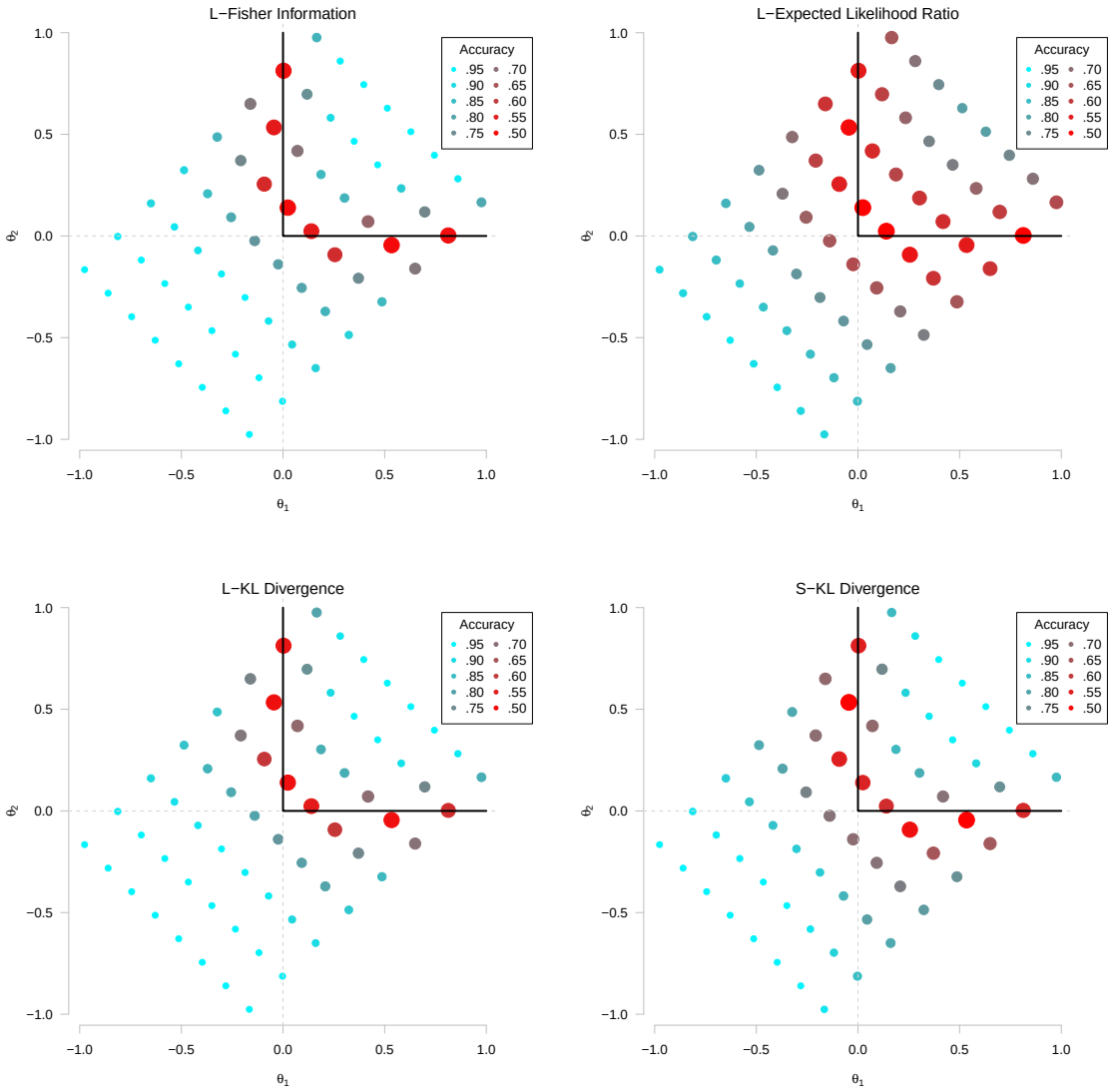

Figure D.9: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the non-compensatory classification bound function and the M-SCSPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

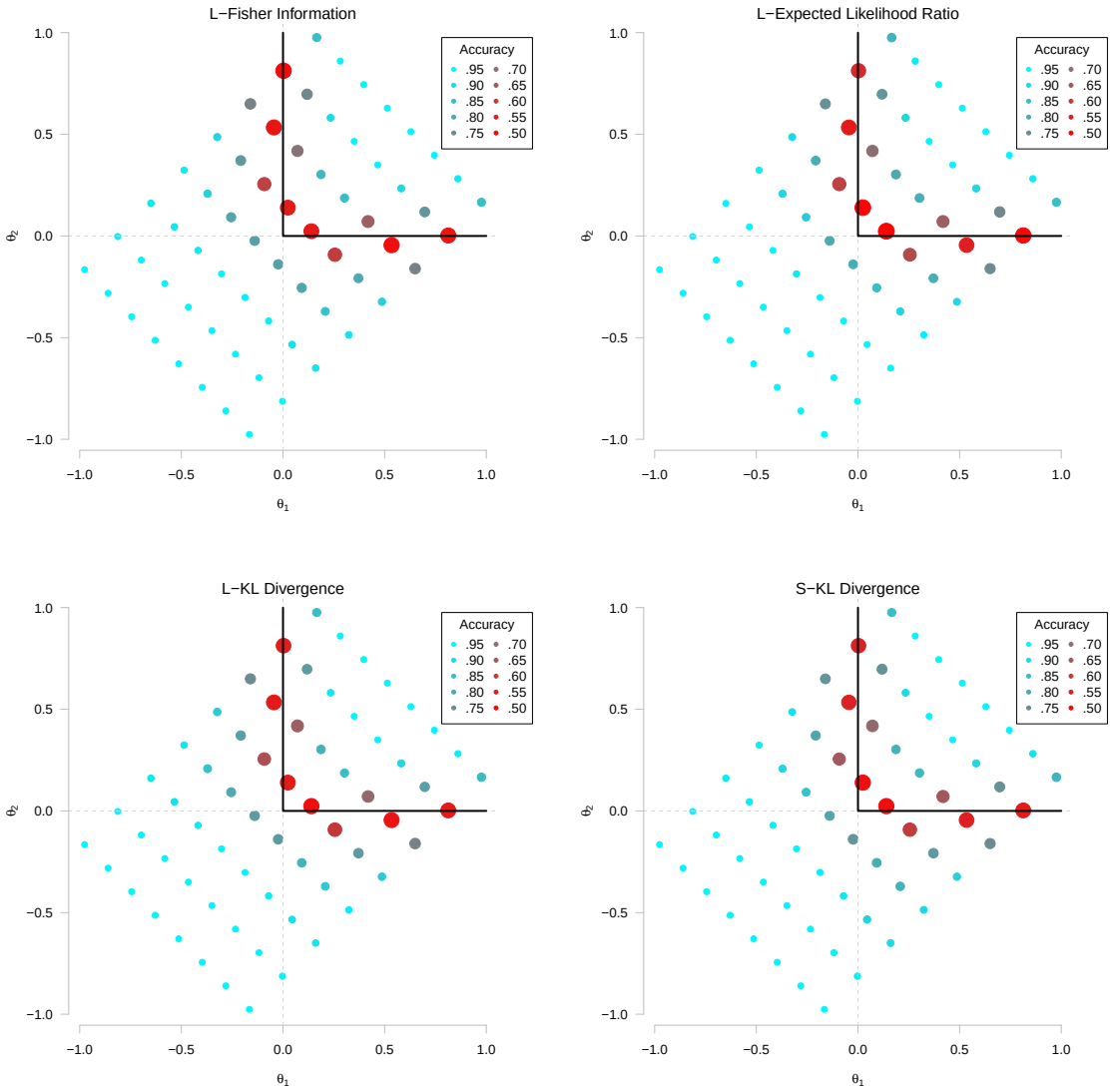

Figure D.10: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the non-compensatory classification bound function and the M-GLR stopping rule with $\delta = .15$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

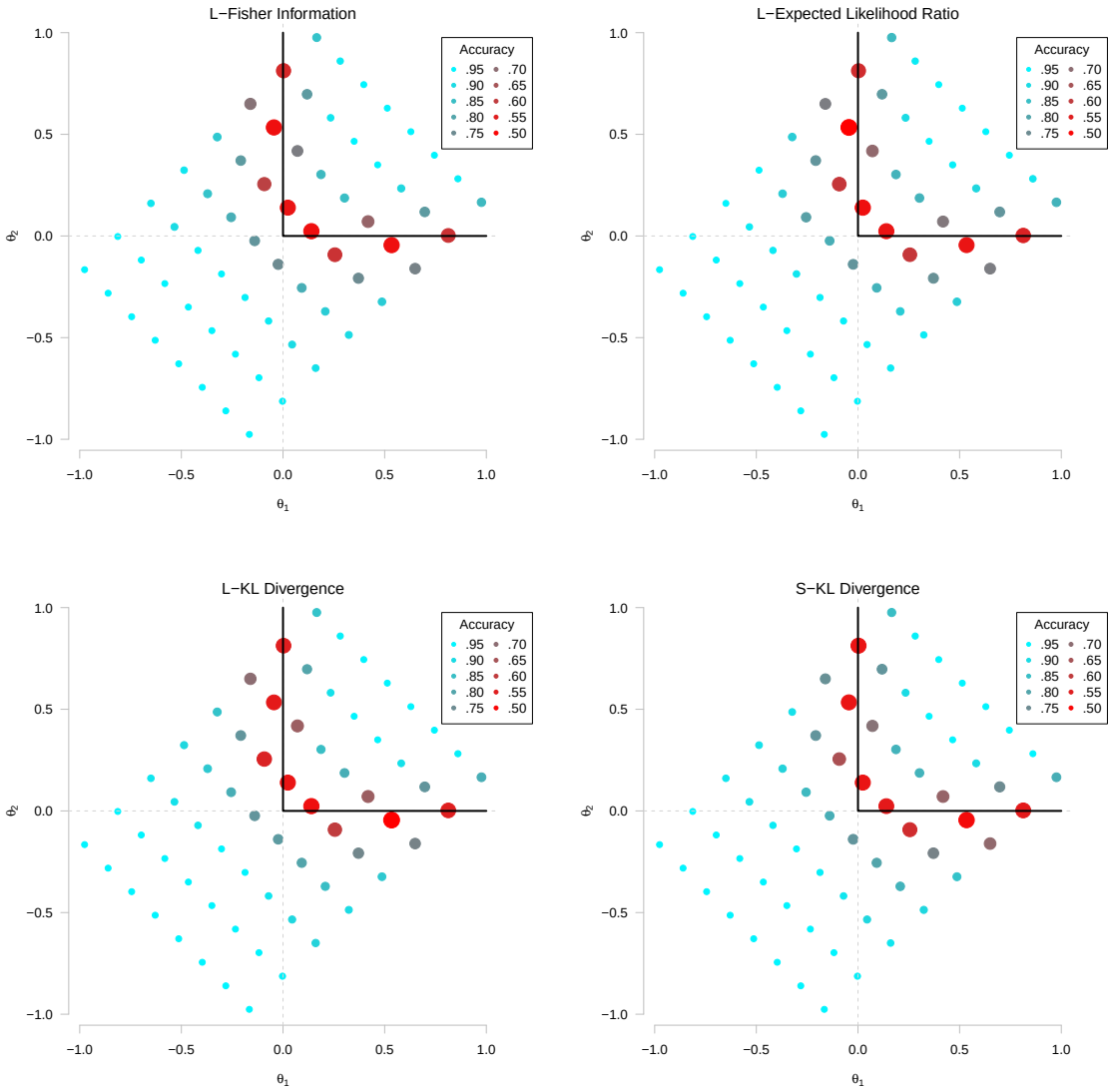

Figure D.11: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the non-compensatory classification bound function and the M-GLR stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

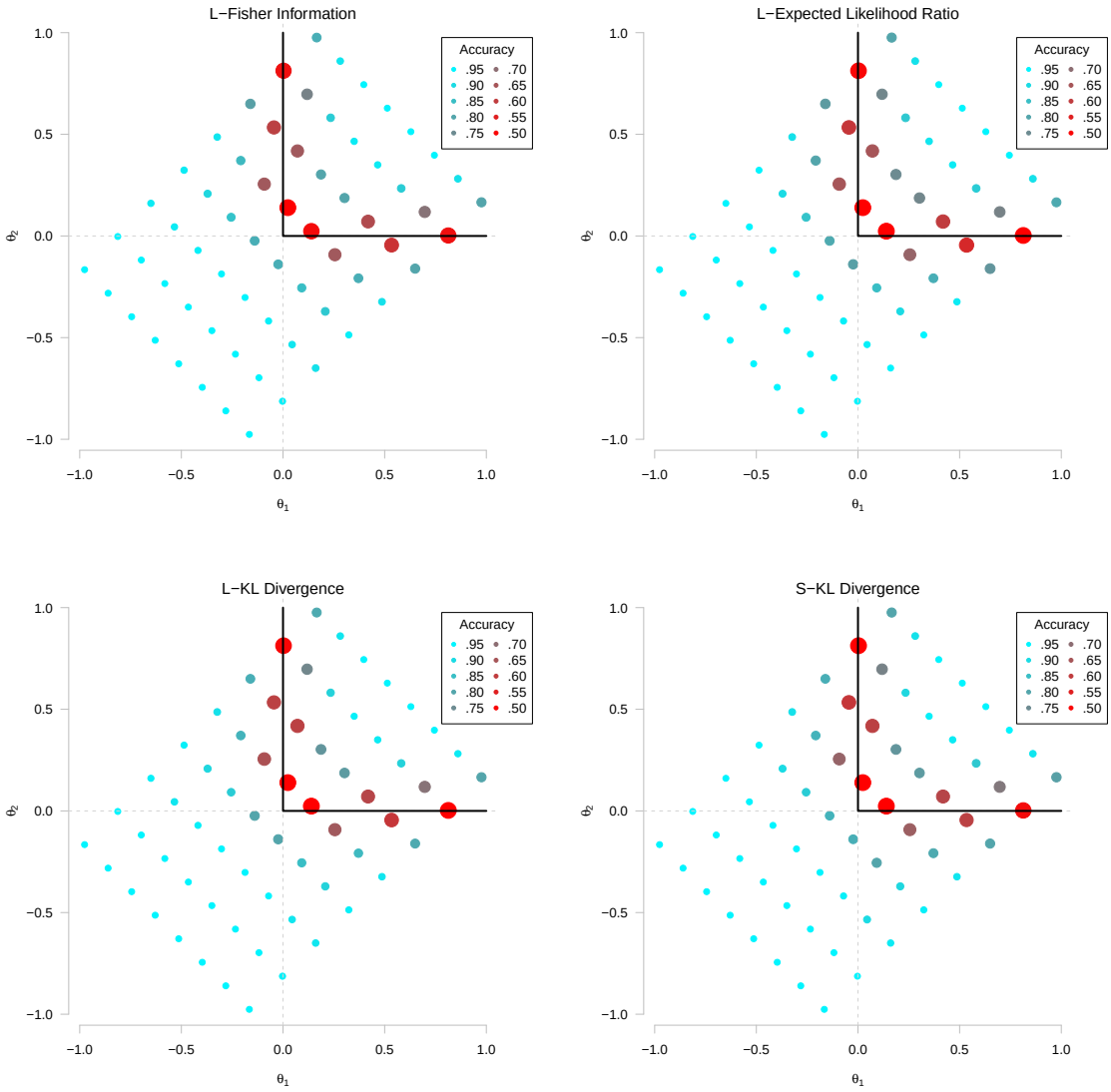

Figure D.12: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the non-compensatory classification bound function and the BCR stopping rule with $\alpha = .05$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

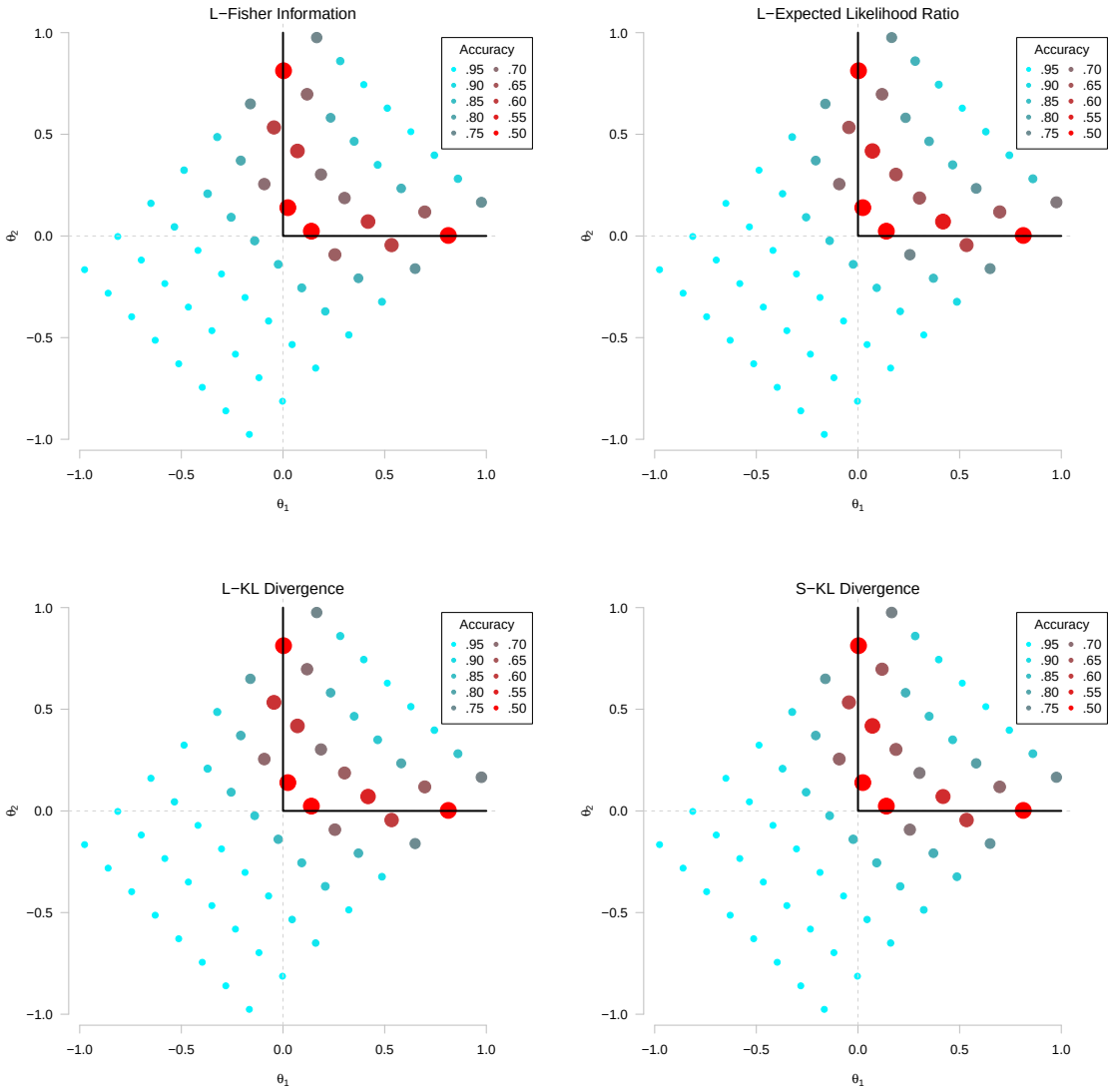

Figure D.13: Scatterplots of the conditional accuracy rate for various vectors of true ability when using the non-compensatory classification bound function and the BCR stopping rule with $\alpha = .10$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to accuracy rate. See the left-most panel of Figure D.1 for more information.

D.2 Test Length Plots

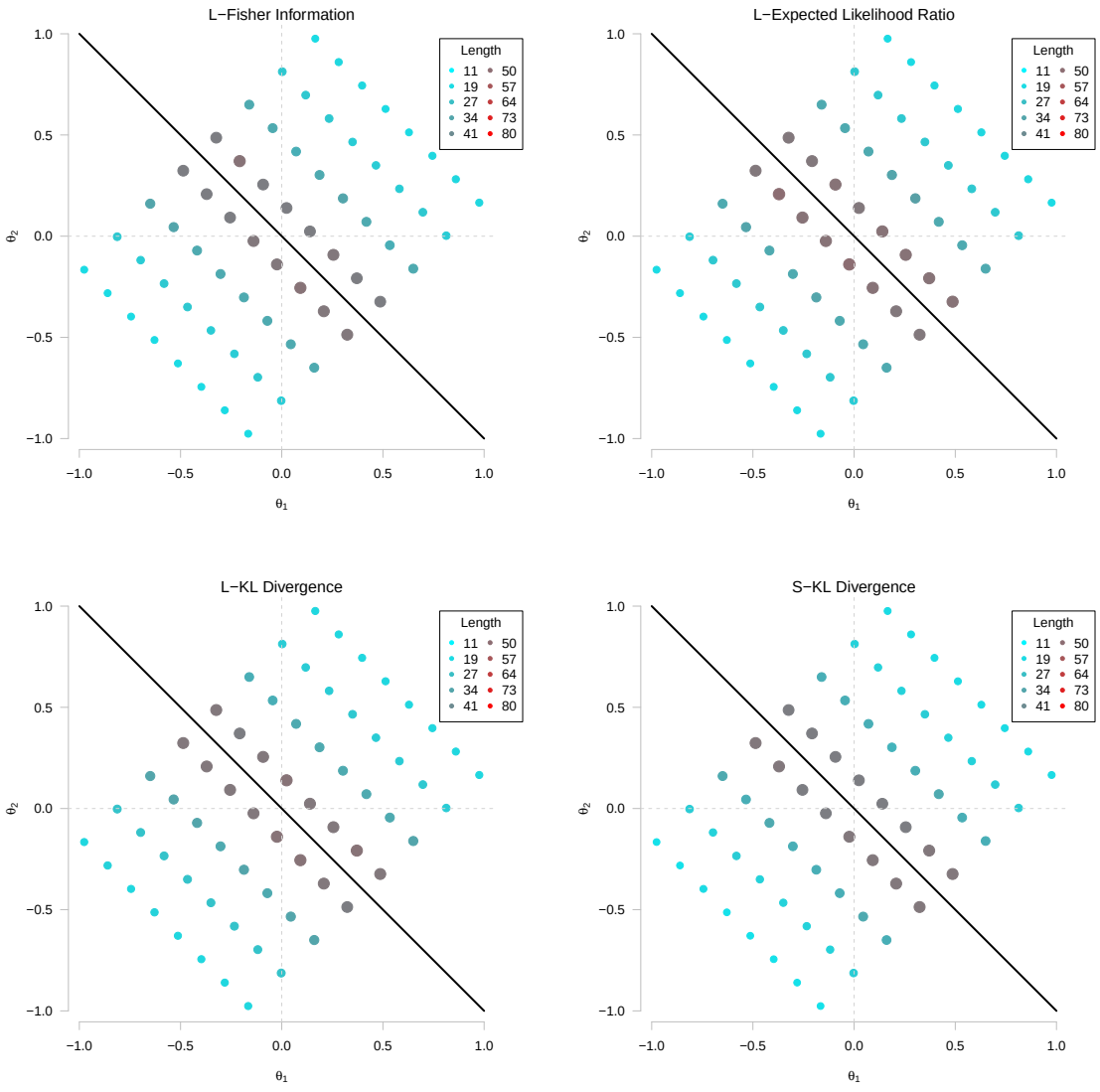

Figure D.14: Scatterplots of the conditional average test length for various vectors of true ability when using the compensatory classification bound function and the C-SPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

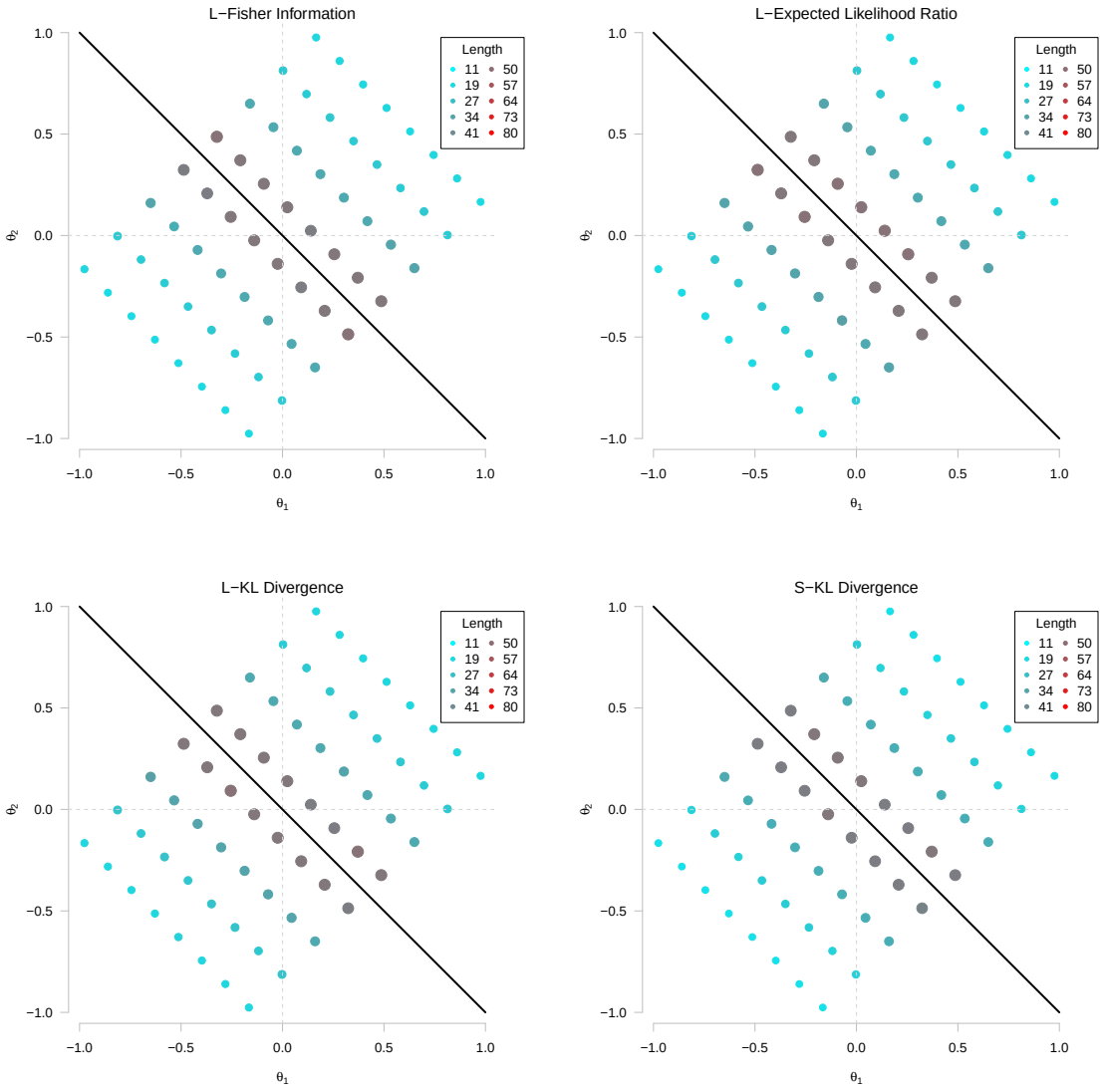

Figure D.15: Scatterplots of the conditional average test length for various vectors of true ability when using the compensatory classification bound function and the M-SCSPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

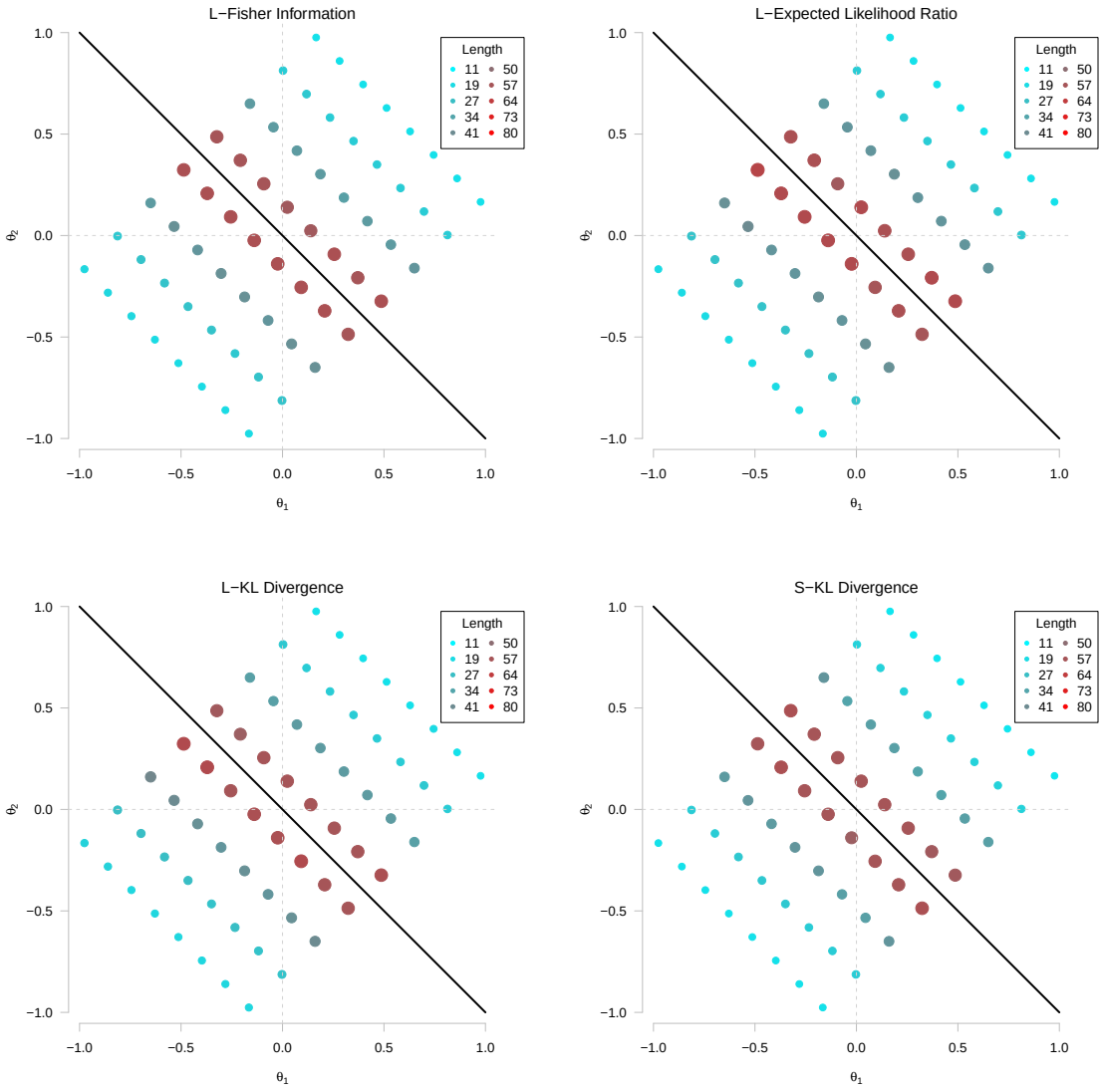

Figure D.16: Scatterplots of the conditional average test length for various vectors of true ability when using the compensatory classification bound function and the M-GLR stopping rule with $\delta = .15$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

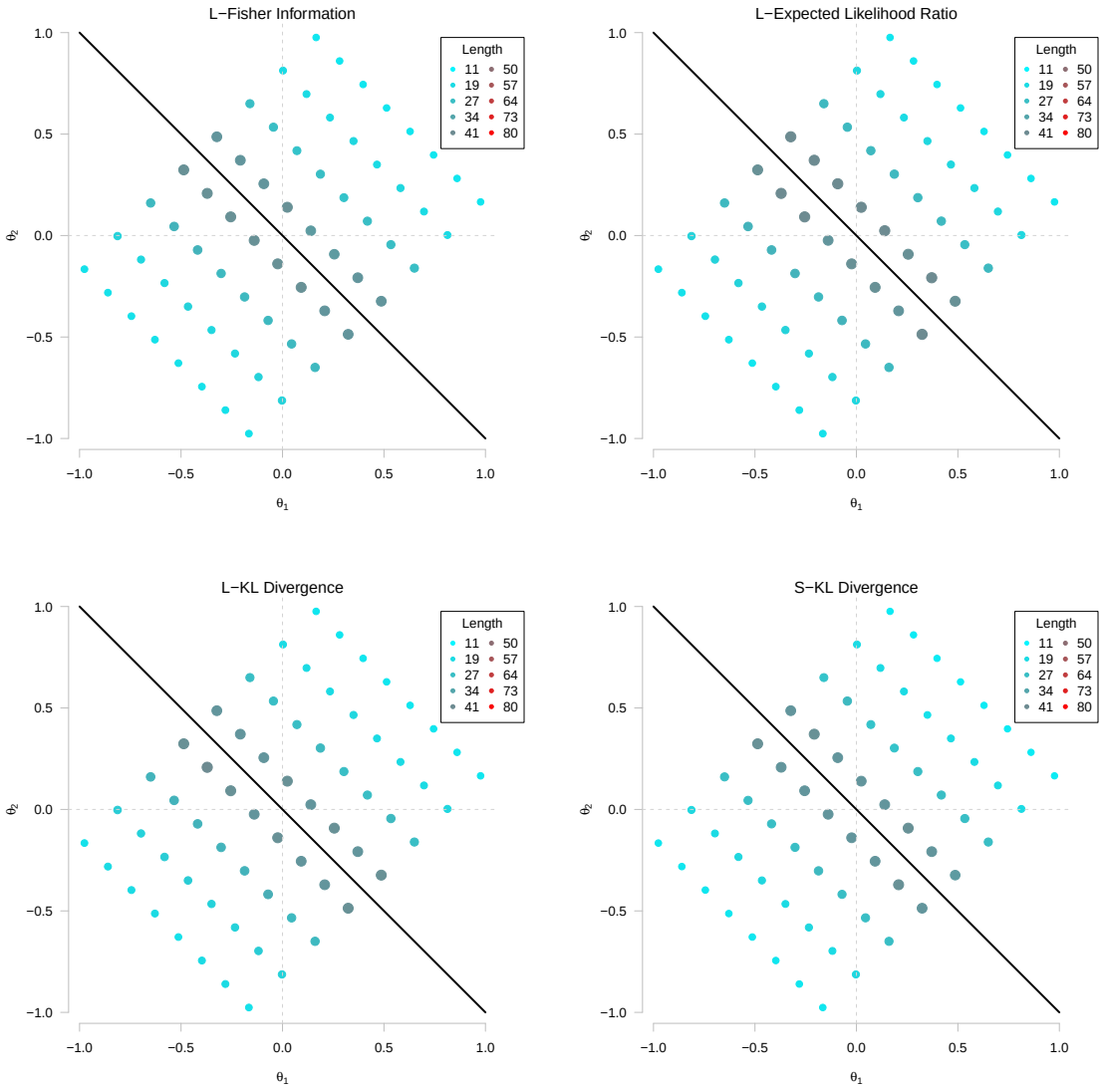

Figure D.17: Scatterplots of the conditional average test length for various vectors of true ability when using the compensatory classification bound function and the M-GLR stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

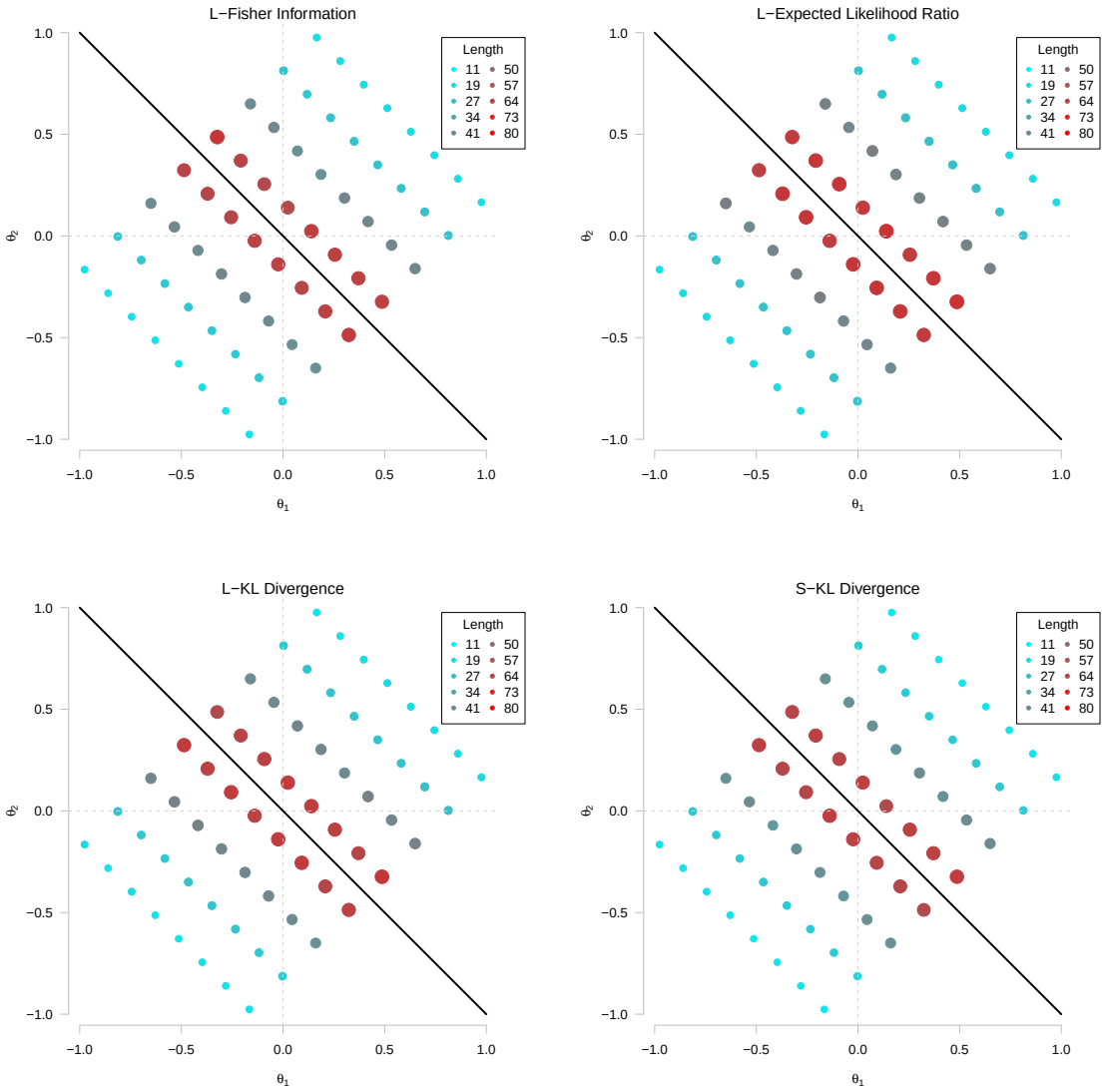

Figure D.18: Scatterplots of the conditional average test length for various vectors of true ability when using the compensatory classification bound function and the BCR stopping rule with $\alpha = .05$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

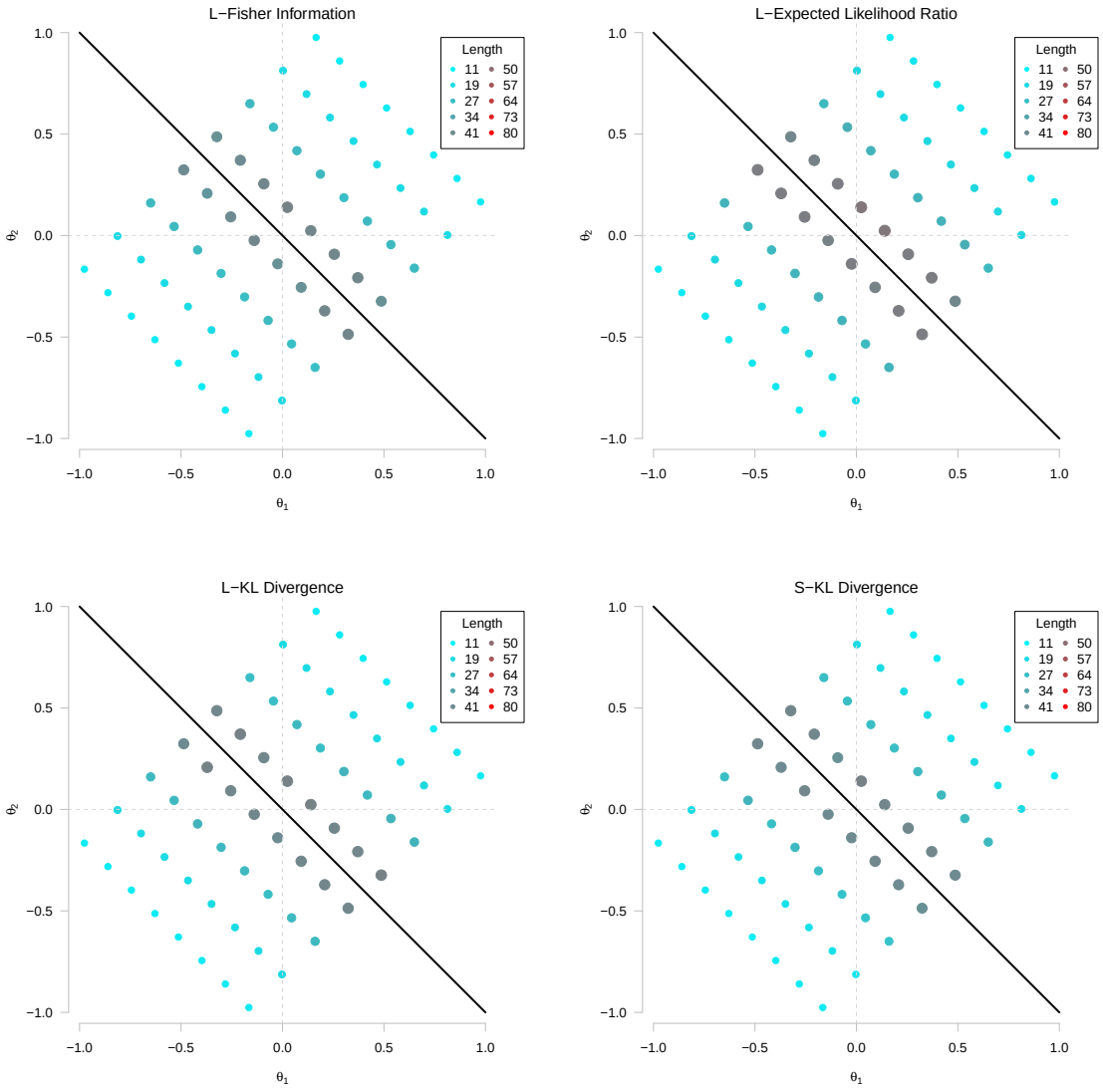

Figure D.19: Scatterplots of the conditional average test length for various vectors of true ability when using the compensatory classification bound function and the BCR stopping rule with $\alpha = .10$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

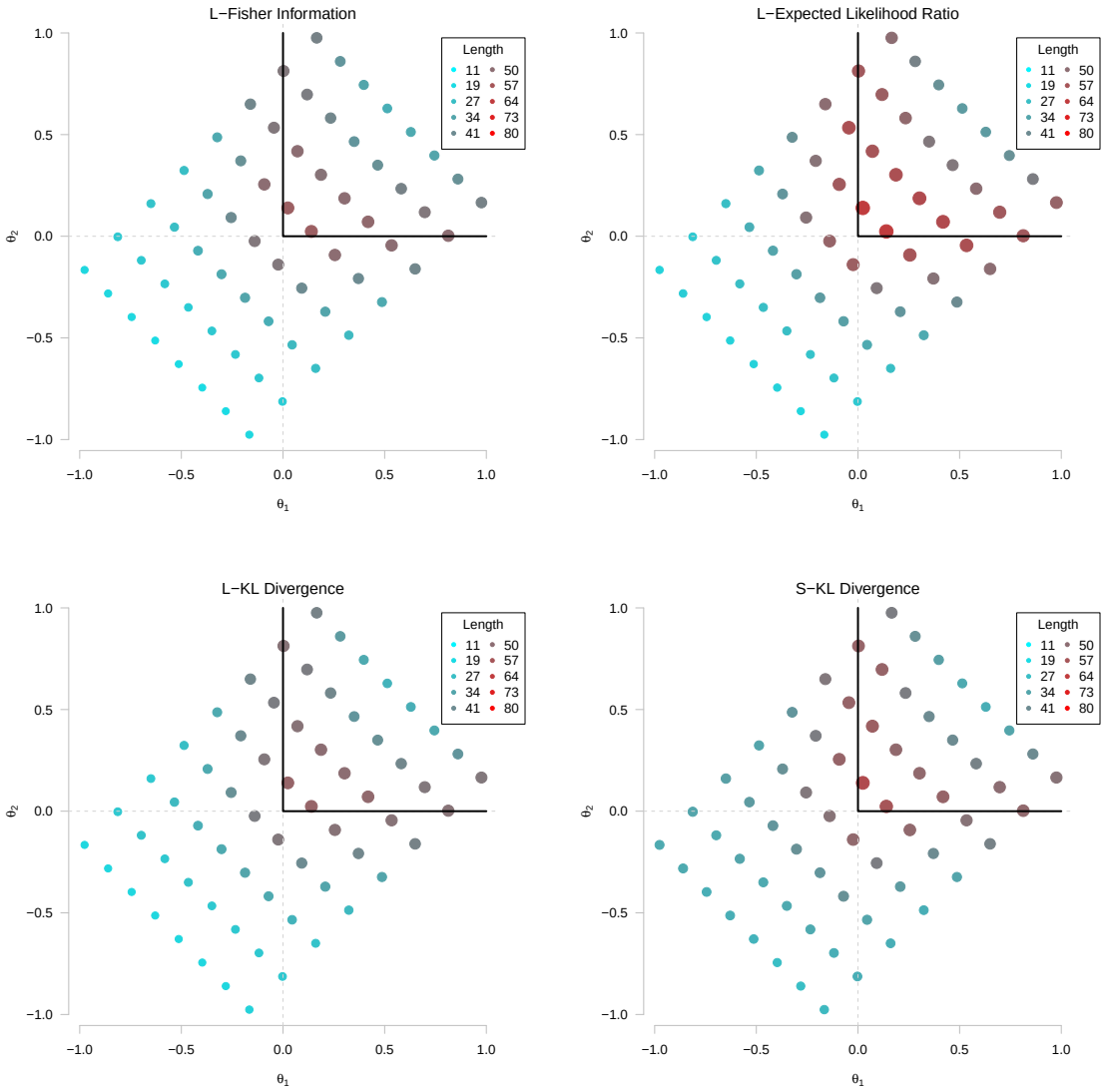

Figure D.20: Scatterplots of the conditional average test length for various vectors of true ability when using the non-compensatory classification bound function and the C-SPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

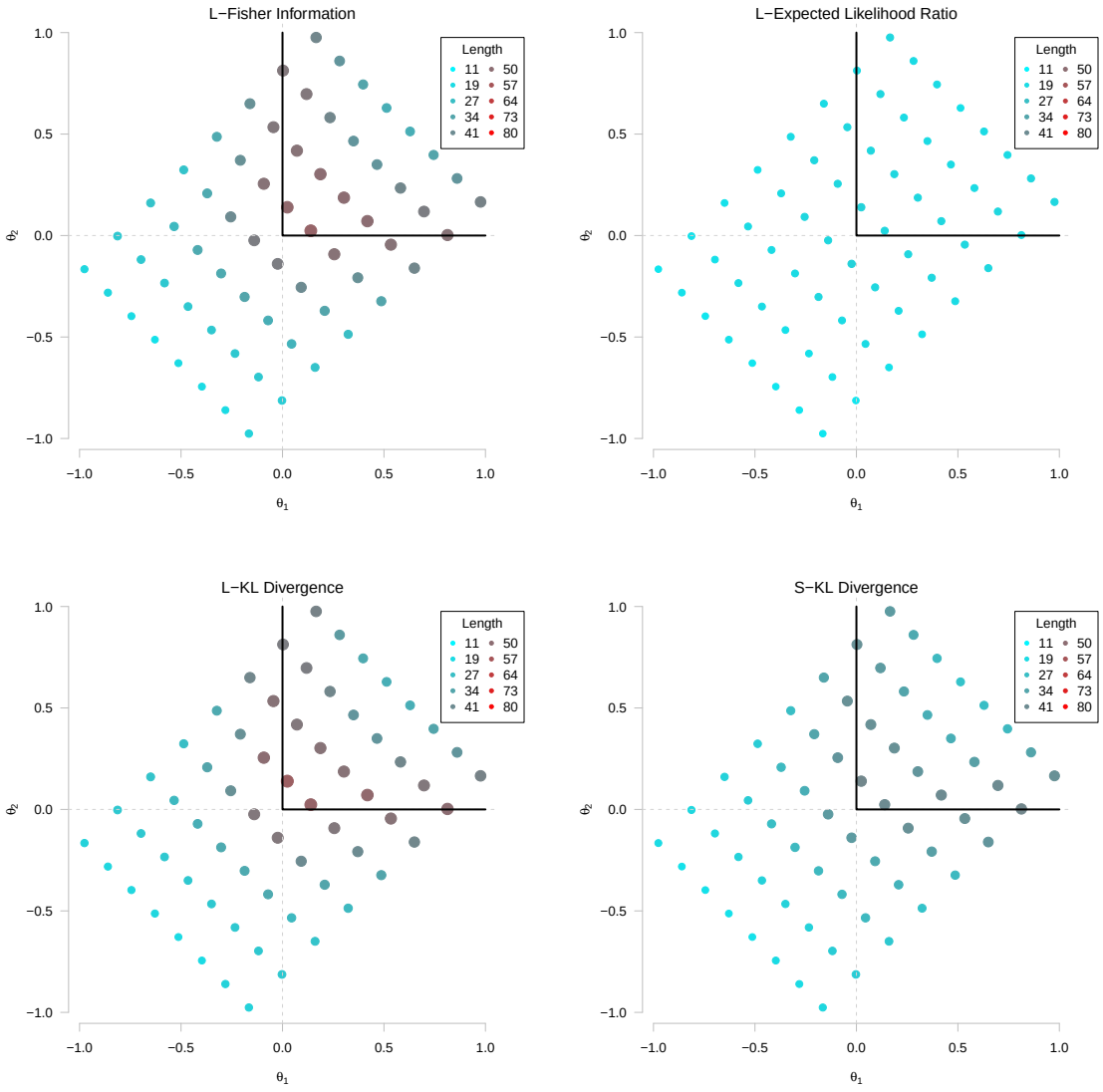

Figure D.21: Scatterplots of the conditional average test length for various vectors of true ability when using the non-compensatory classification bound function and the M-SCSPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

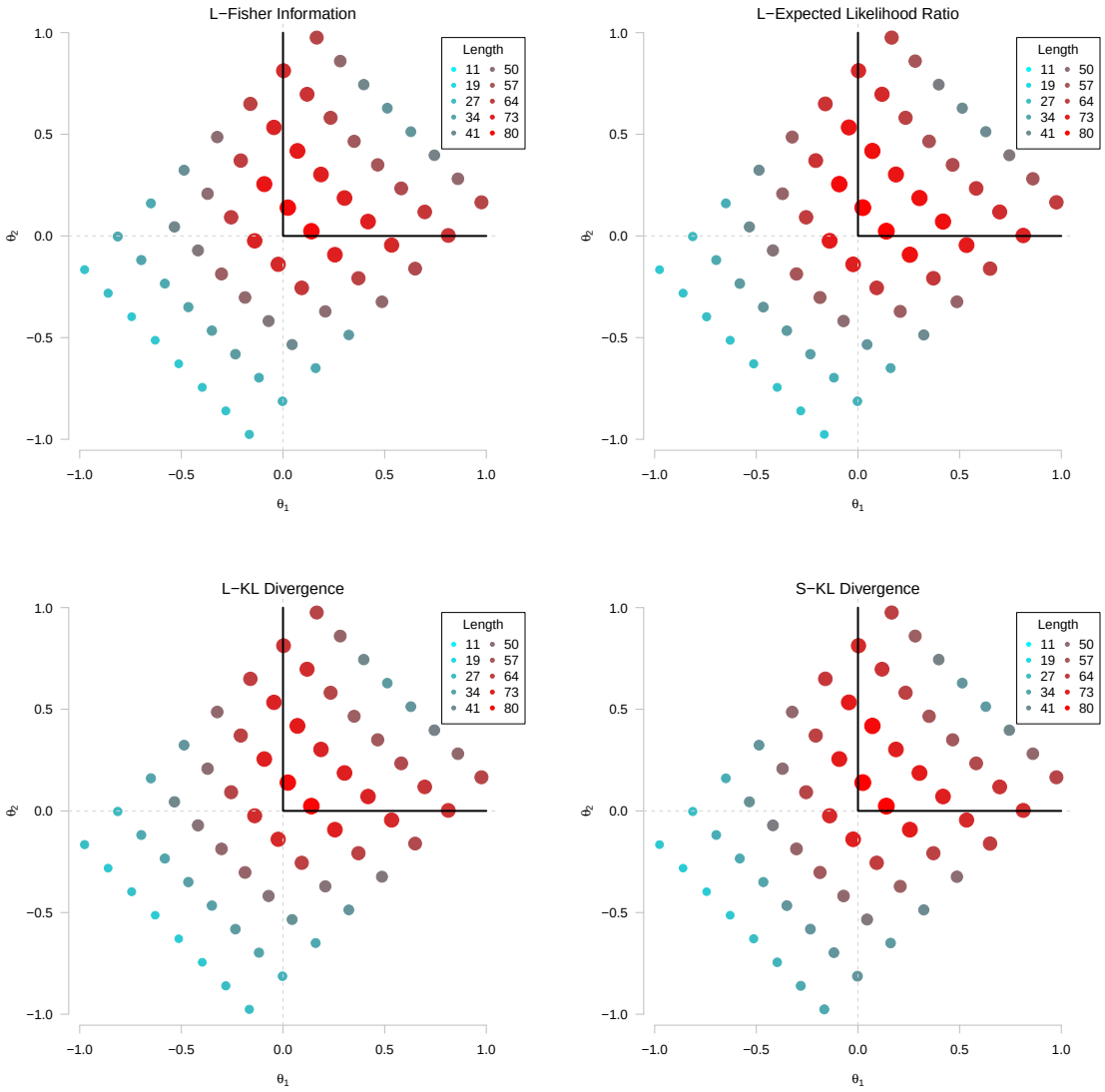

Figure D.22: Scatterplots of the conditional average test length for various vectors of true ability when using the non-compensatory classification bound function and the M-GLR stopping rule with $\delta = .15$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

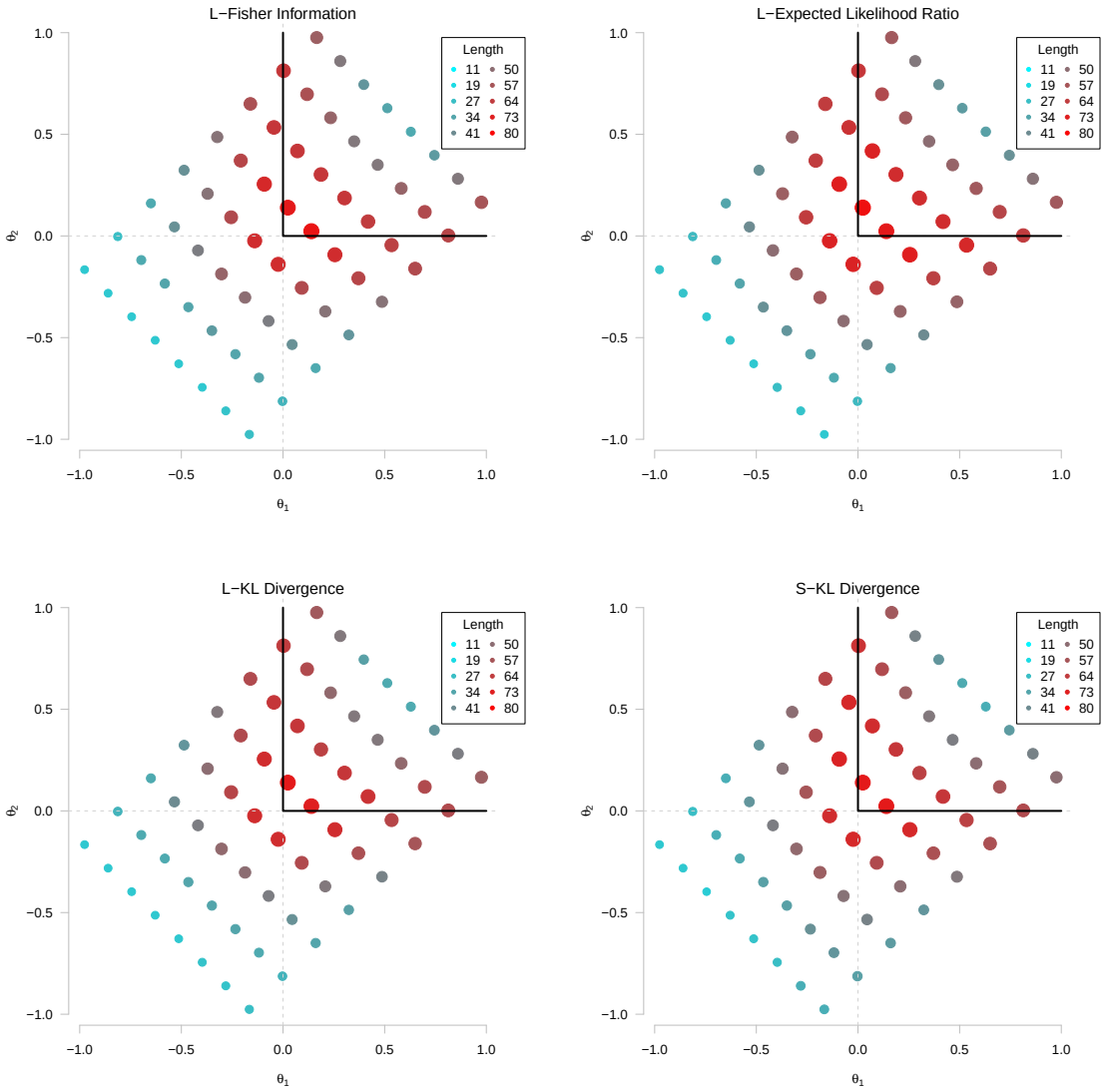

Figure D.23: Scatterplots of the conditional average test length for various vectors of true ability when using the non-compensatory classification bound function and the M-GLR stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

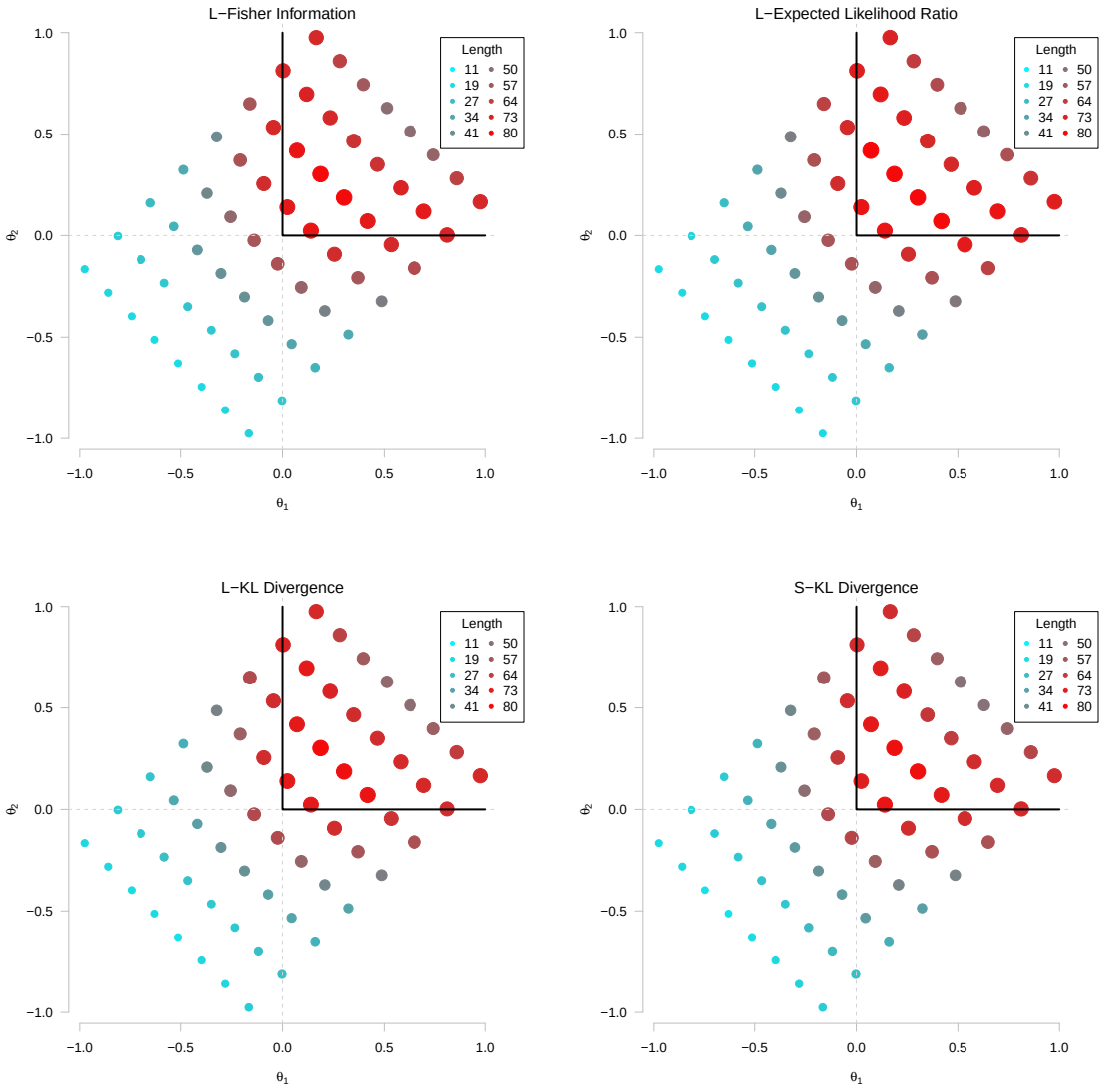

Figure D.24: Scatterplots of the conditional average test length for various vectors of true ability when using the non-compensatory classification bound function and the BCR stopping rule with $\alpha = .05$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

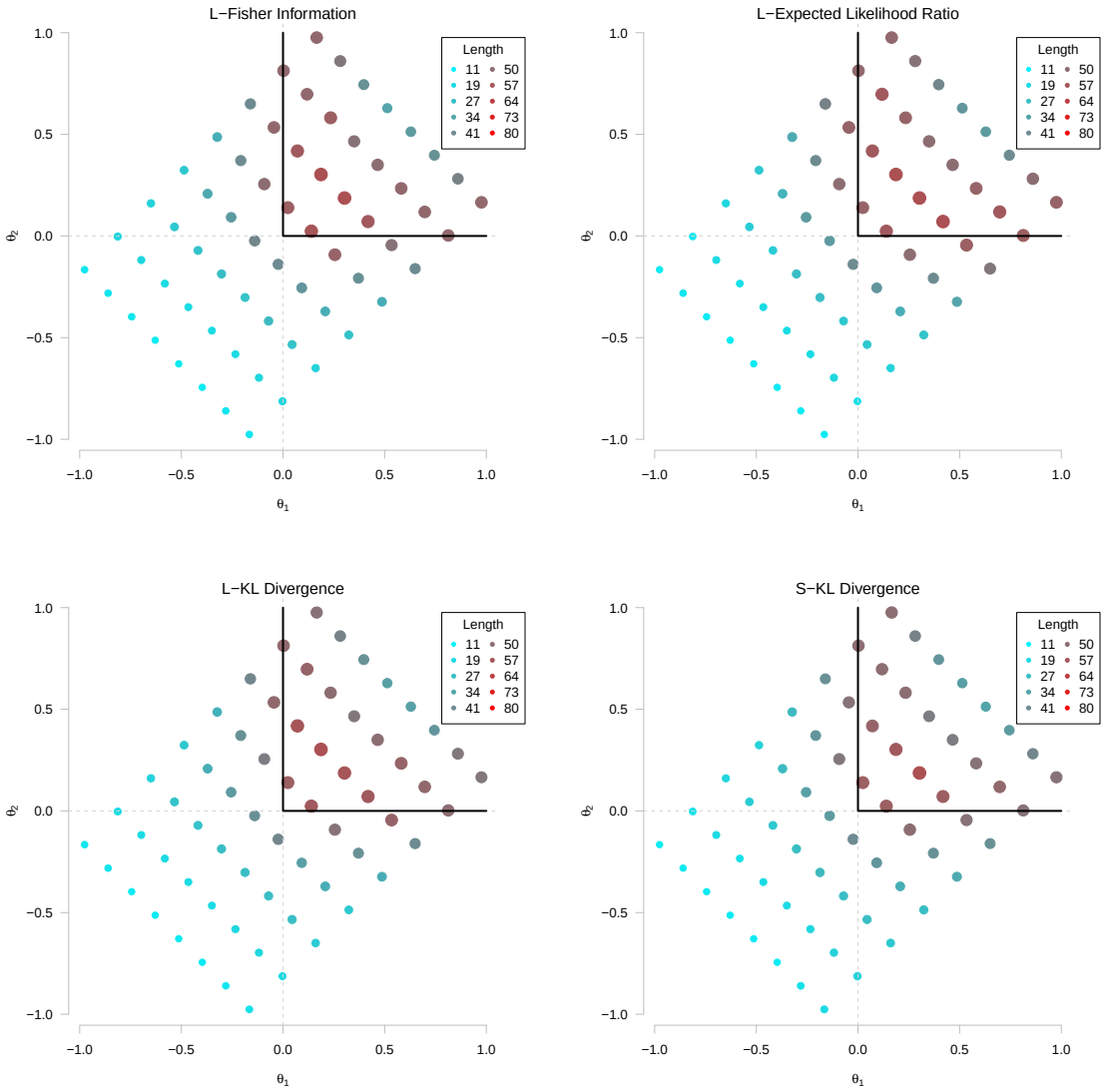

Figure D.25: Scatterplots of the conditional average test length for various vectors of true ability when using the non-compensatory classification bound function and the BCR stopping rule with $\alpha = .10$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

D.3 Loss Plots

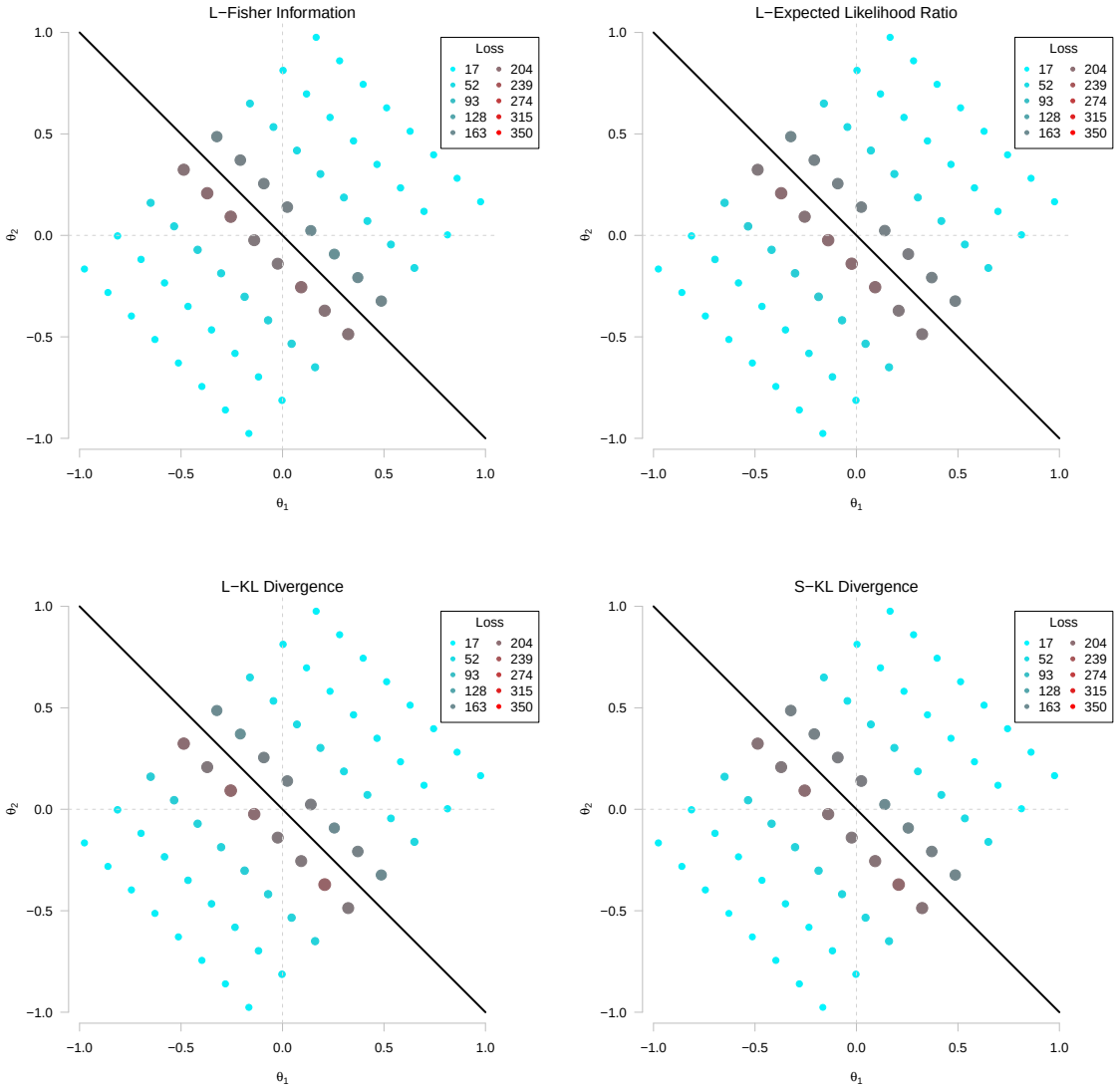

Figure D.26: Scatterplots of the conditional average loss for various vectors of true ability when using the compensatory classification bound function and the C-SPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to loss. See the right-most panel of Figure D.1 for more information.

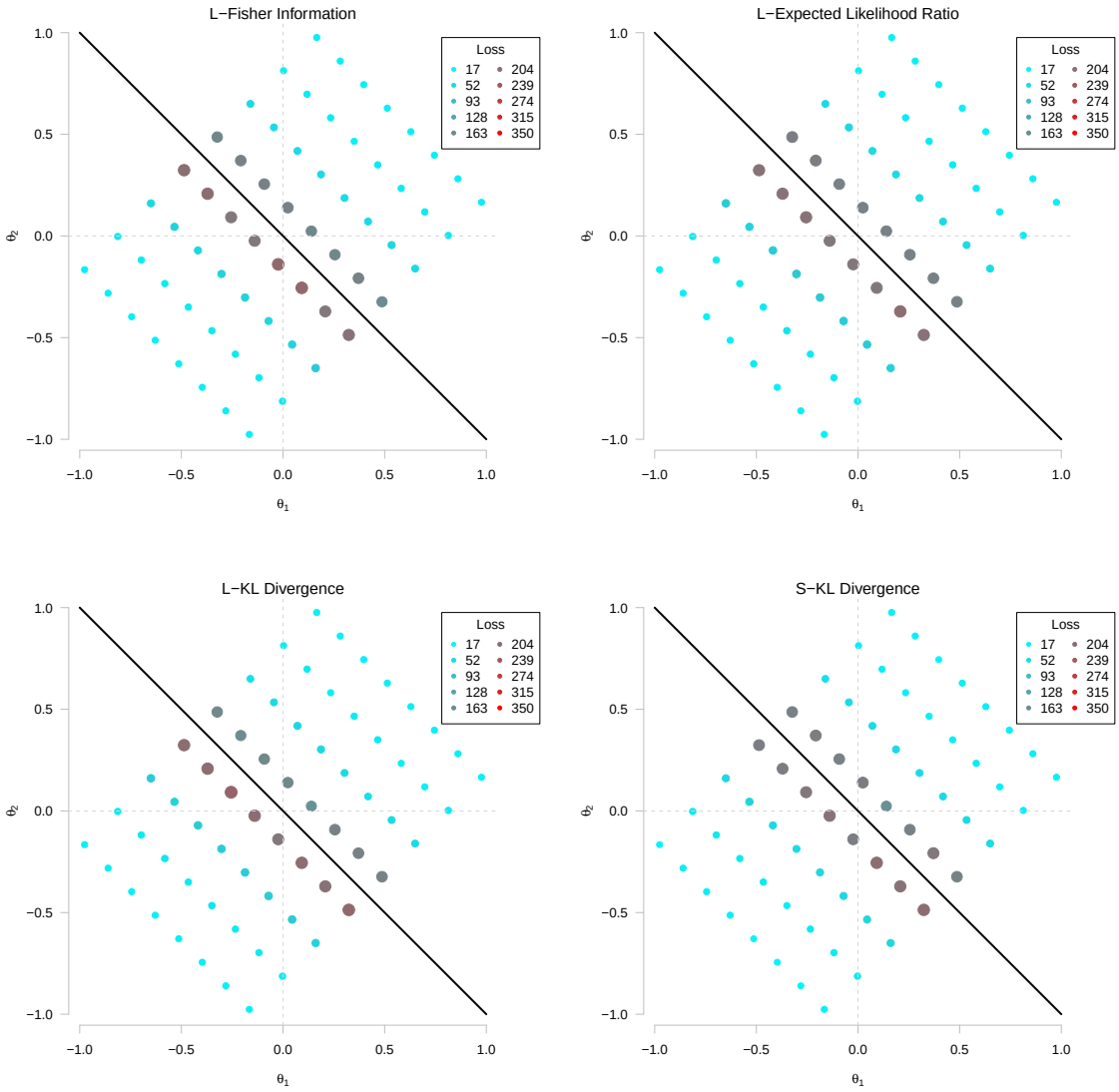

Figure D.27: Scatterplots of the conditional average loss for various vectors of true ability when using the compensatory classification bound function and the M-SCSPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to loss. See the right-most panel of Figure D.1 for more information.

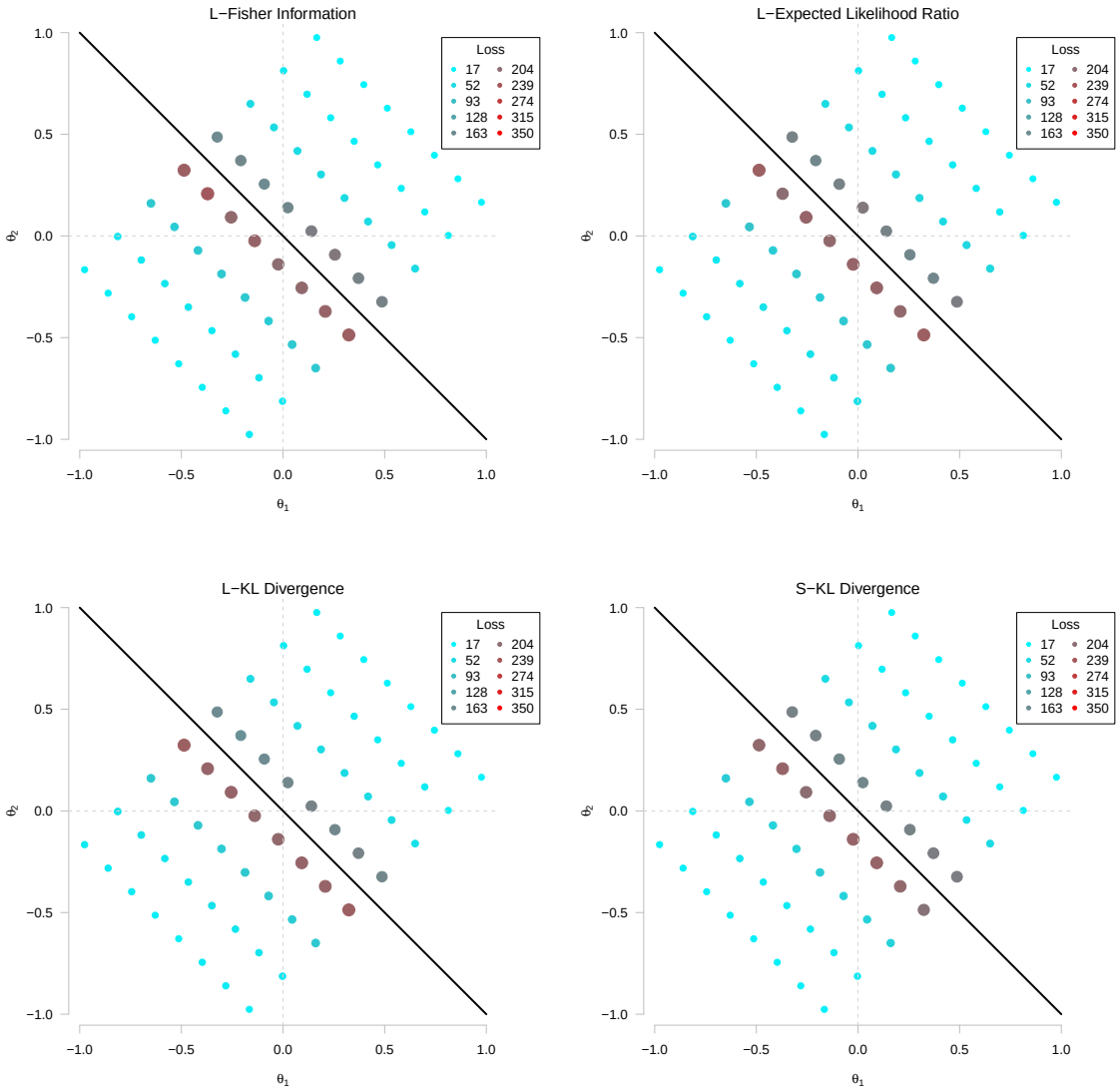

Figure D.28: Scatterplots of the conditional average loss for various vectors of true ability when using the compensatory classification bound function and the M-GLR stopping rule with $\delta = .15$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to loss. See the right-most panel of Figure D.1 for more information.

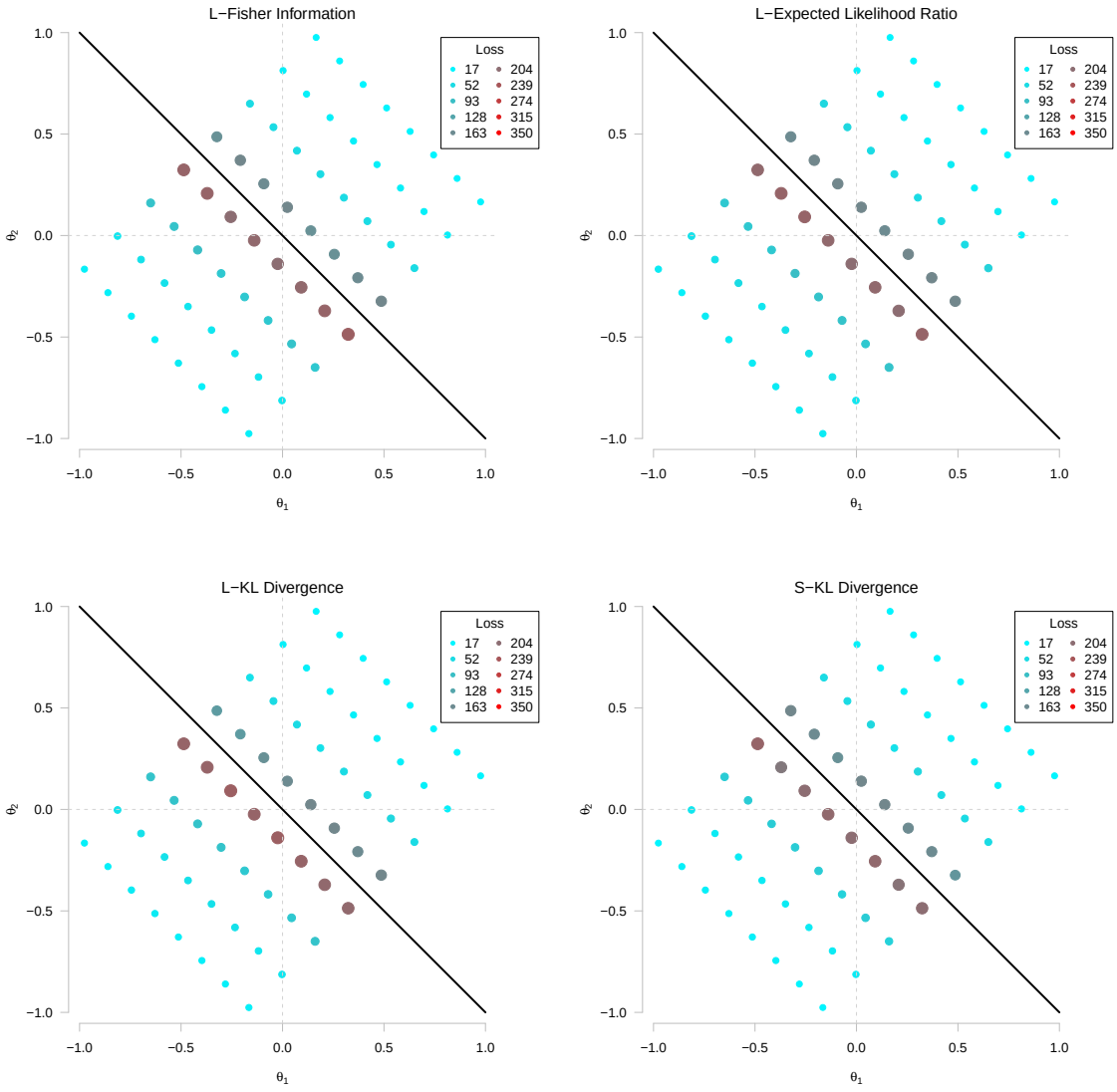

Figure D.29: Scatterplots of the conditional average loss for various vectors of true ability when using the compensatory classification bound function and the M-GLR stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to loss. See the right-most panel of Figure D.1 for more information.

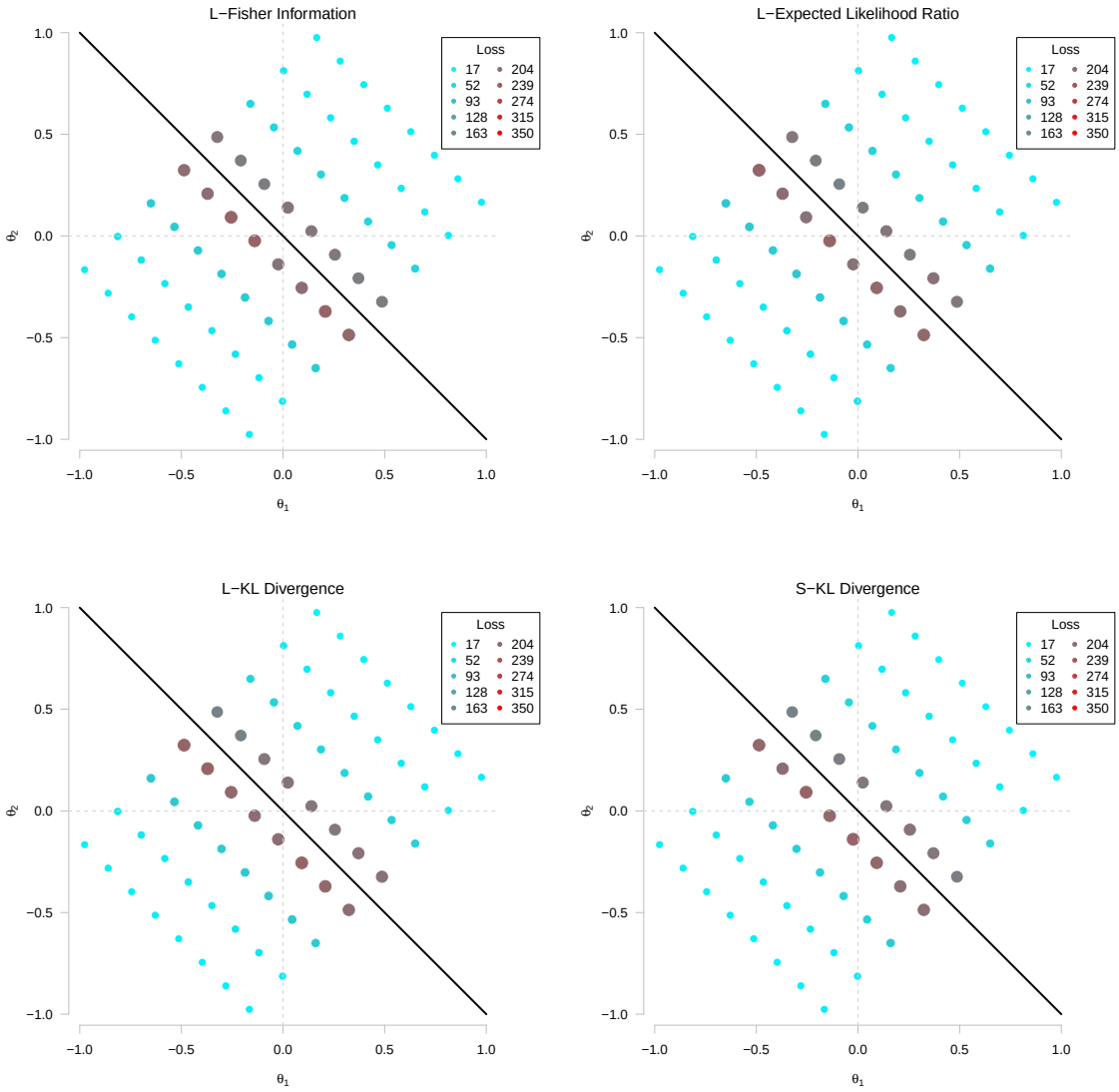

Figure D.30: Scatterplots of the conditional average loss for various vectors of true ability when using the compensatory classification bound function and the BCR stopping rule with $\alpha = .05$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to loss. See the right-most panel of Figure D.1 for more information.

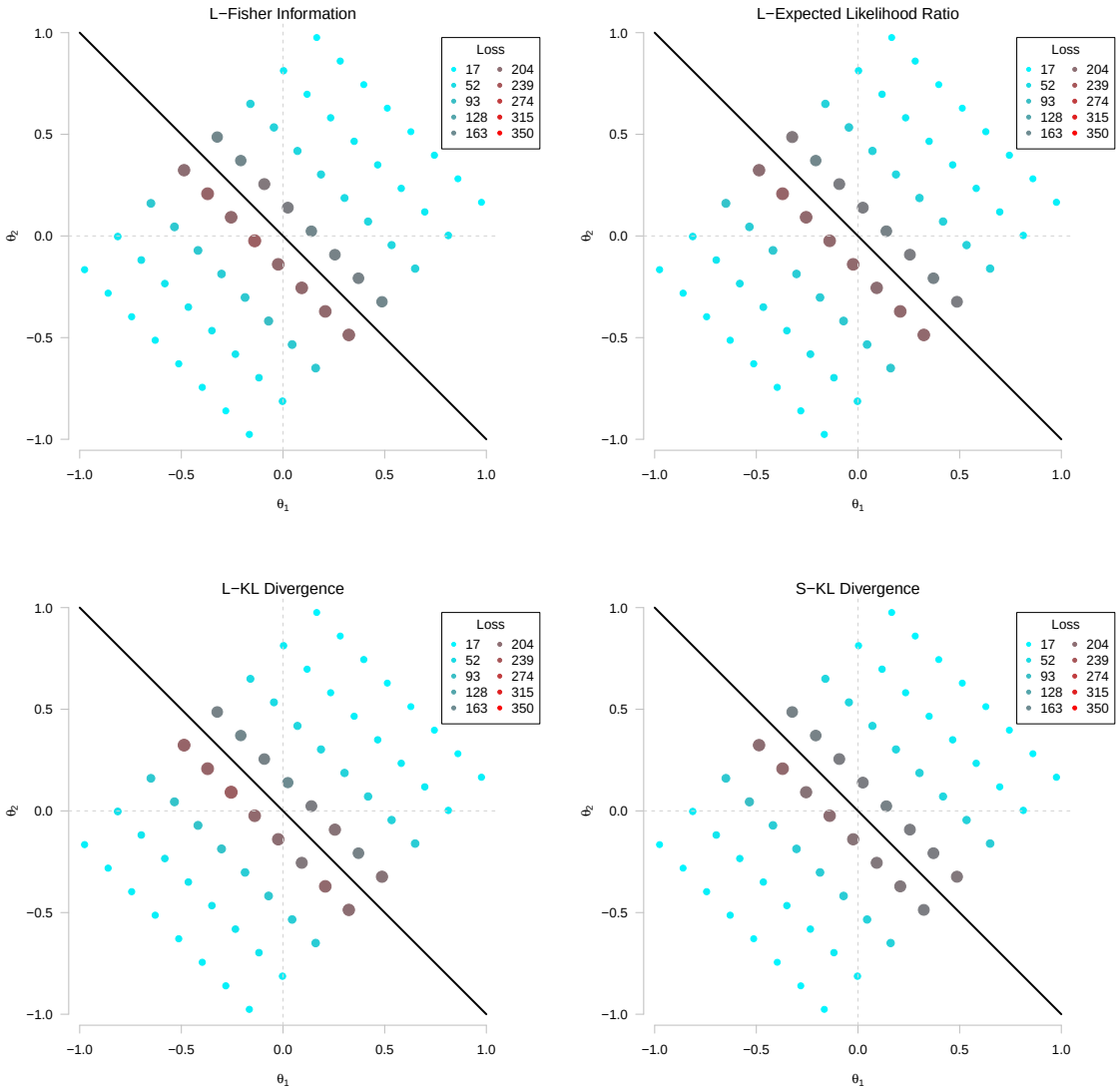

Figure D.31: Scatterplots of the conditional average loss for various vectors of true ability when using the compensatory classification bound function and the BCR stopping rule with $\alpha = .10$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to loss. See the right-most panel of Figure D.1 for more information.

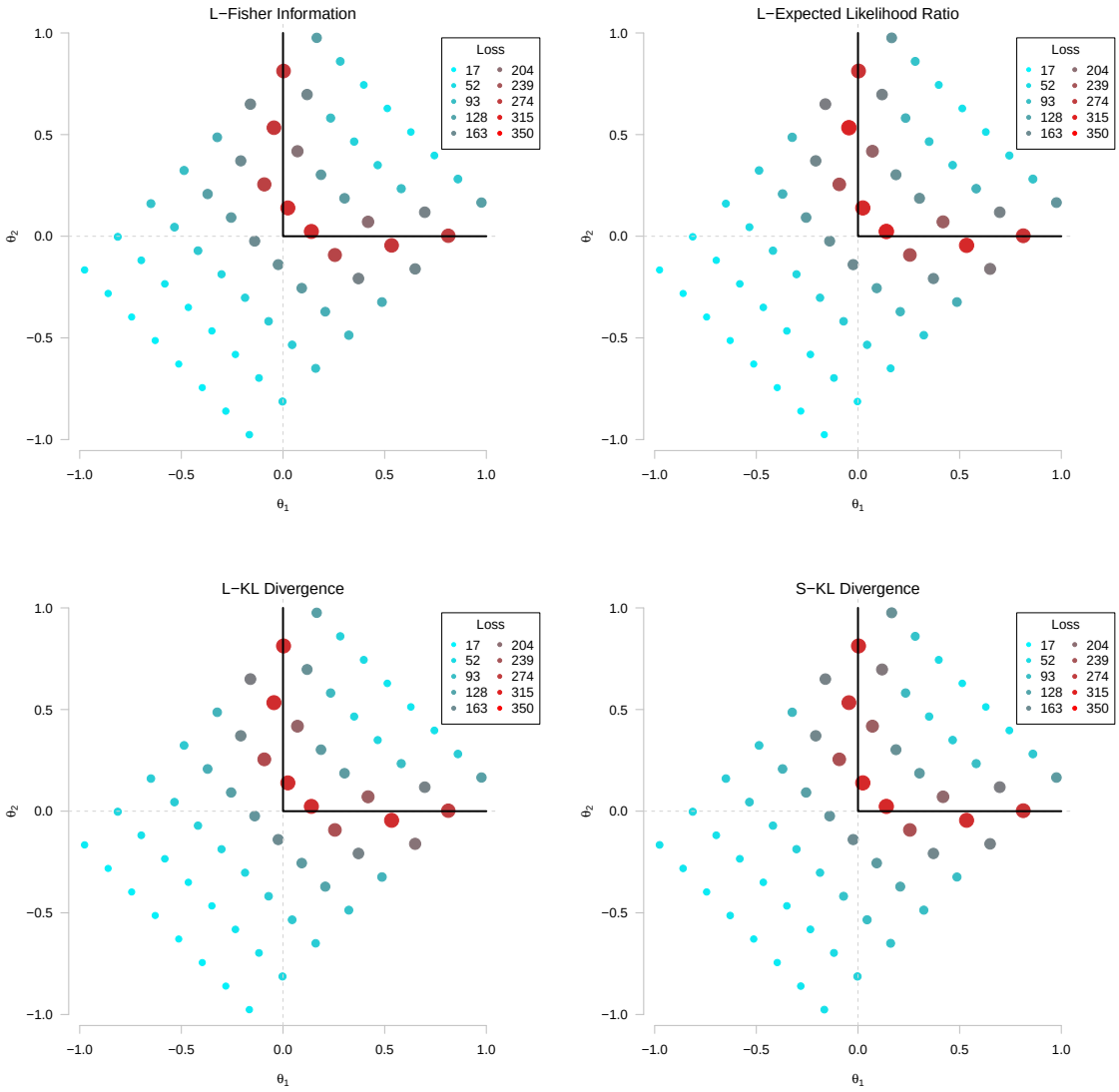

Figure D.32: Scatterplots of the conditional average loss for various vectors of true ability when using the non-compensatory classification bound function and the C-SPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to loss. See the right-most panel of Figure D.1 for more information.

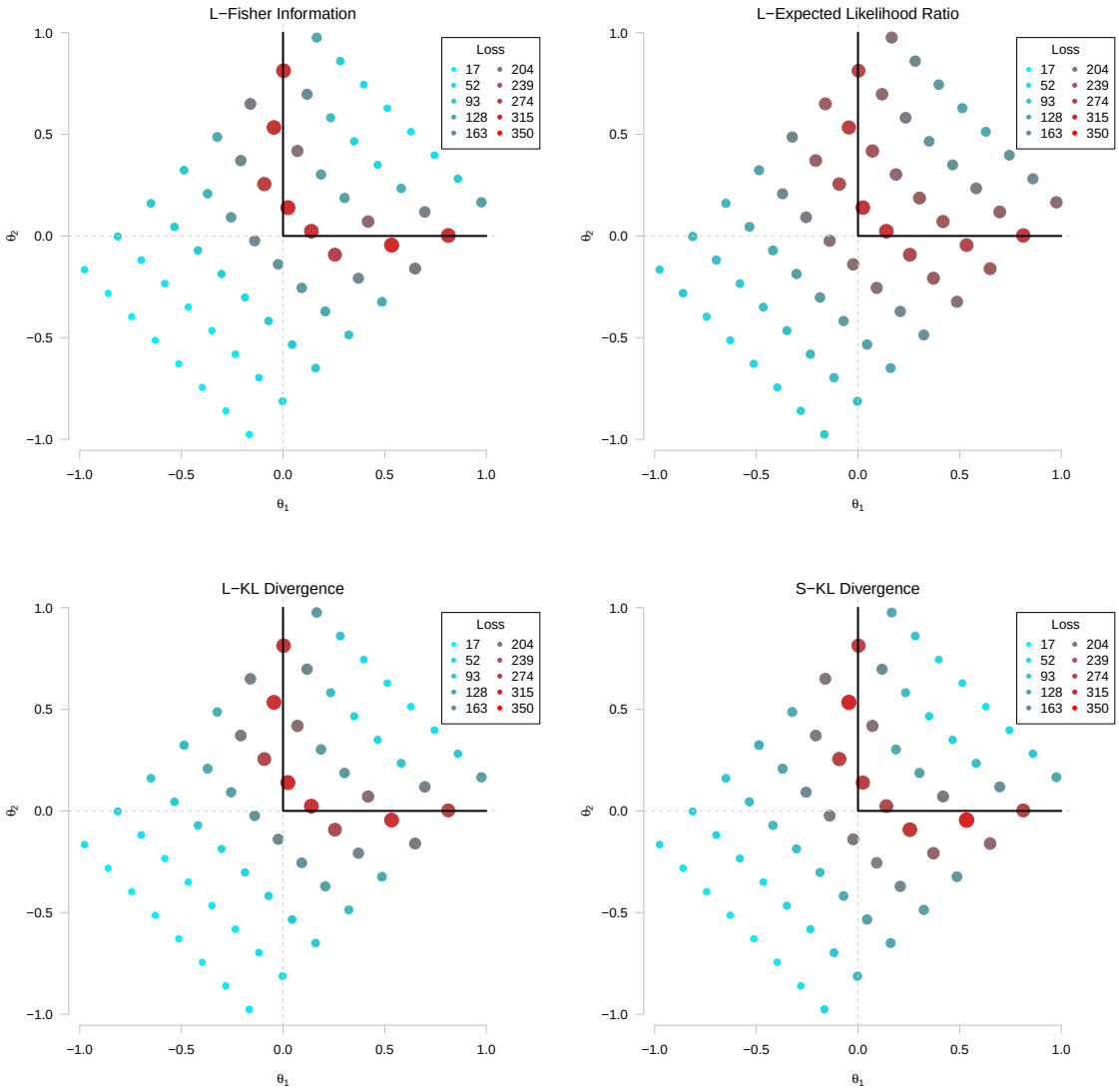

Figure D.33: Scatterplots of the conditional average loss for various vectors of true ability when using the non-compensatory classification bound function and the M-SCSPRT stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to loss. See the right-most panel of Figure D.1 for more information.

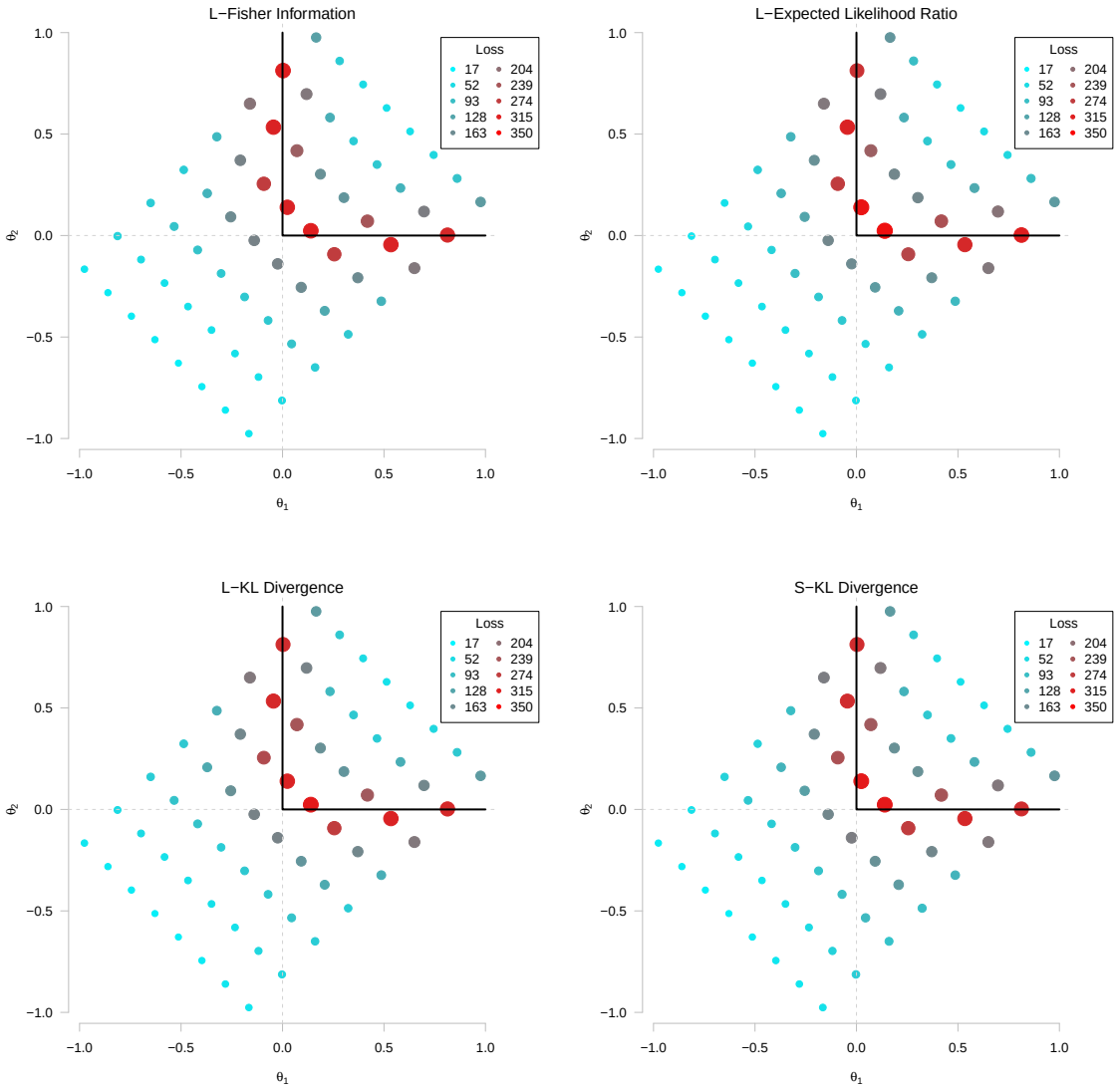

Figure D.34: Scatterplots of the conditional average loss for various vectors of true ability when using the non-compensatory classification bound function and the M-GLR stopping rule with $\delta = .15$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to loss. See the right-most panel of Figure D.1 for more information.

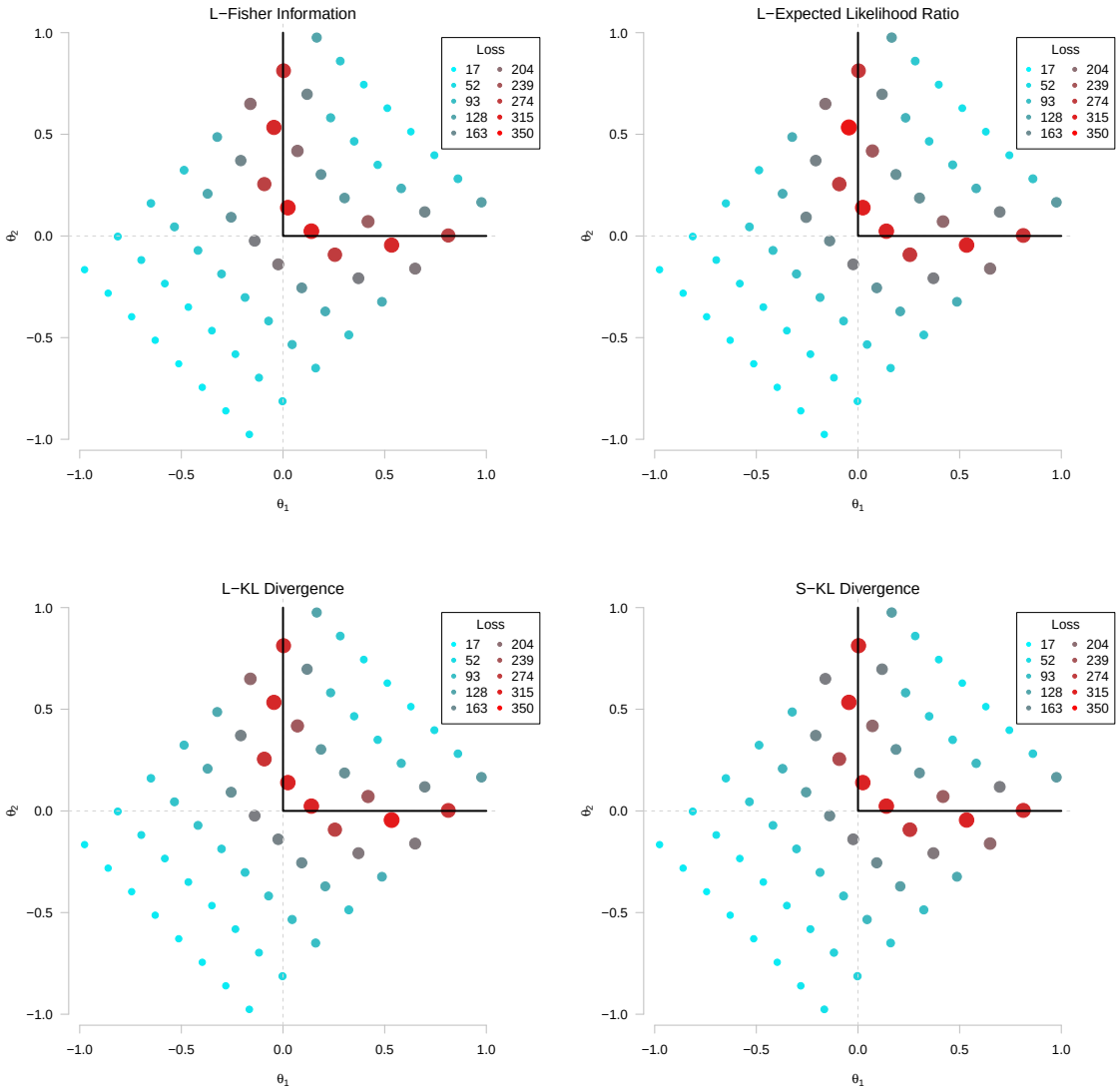

Figure D.35: Scatterplots of the conditional average loss for various vectors of true ability when using the non-compensatory classification bound function and the M-GLR stopping rule with $\delta = .25$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to loss. See the right-most panel of Figure D.1 for more information.

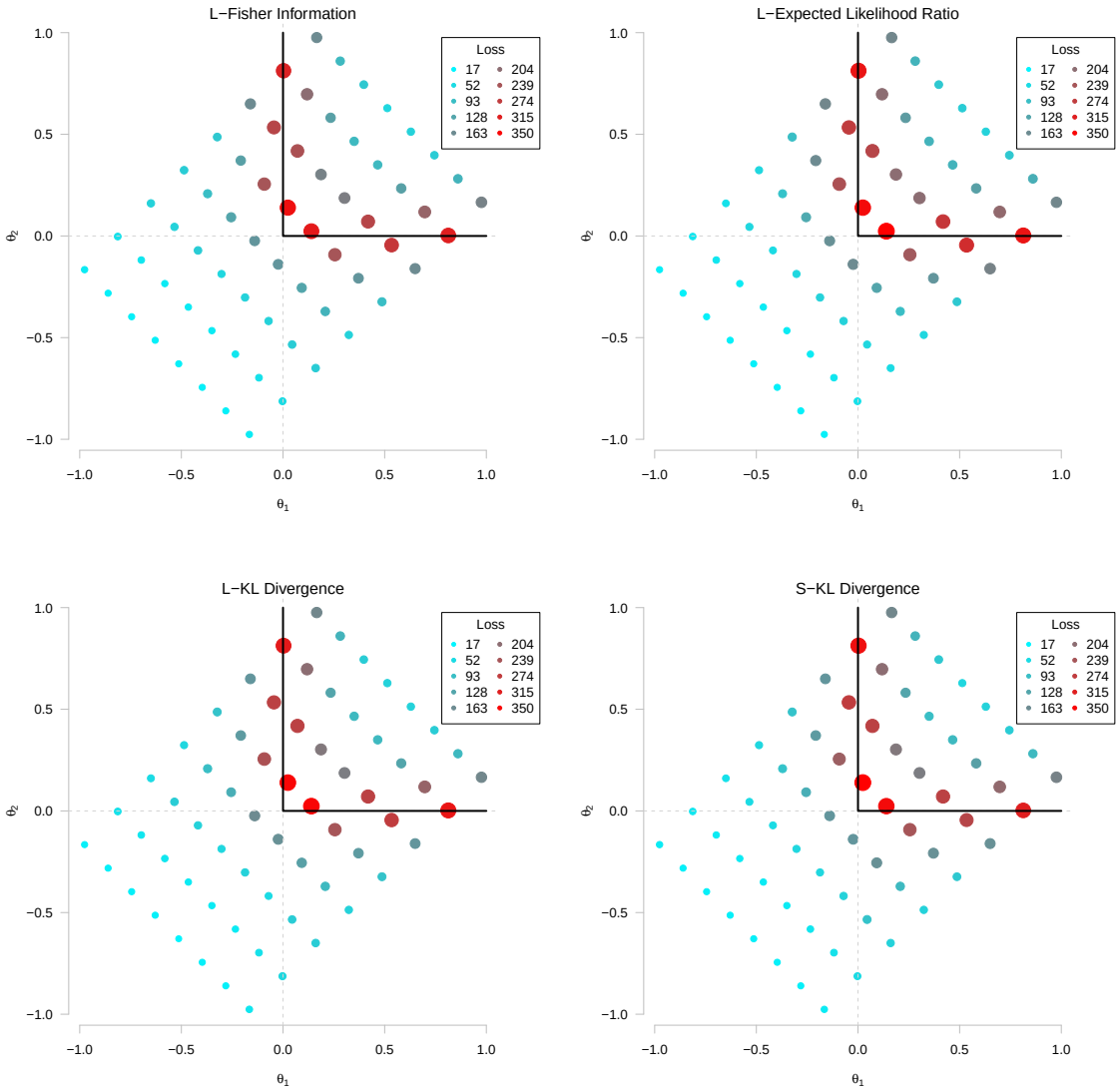

Figure D.36: Scatterplots of the conditional average test length for various vectors of true ability when using the non-compensatory classification bound function and the BCR stopping rule with $\alpha = .05$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to test length. See the middle panel of Figure D.1 for more information.

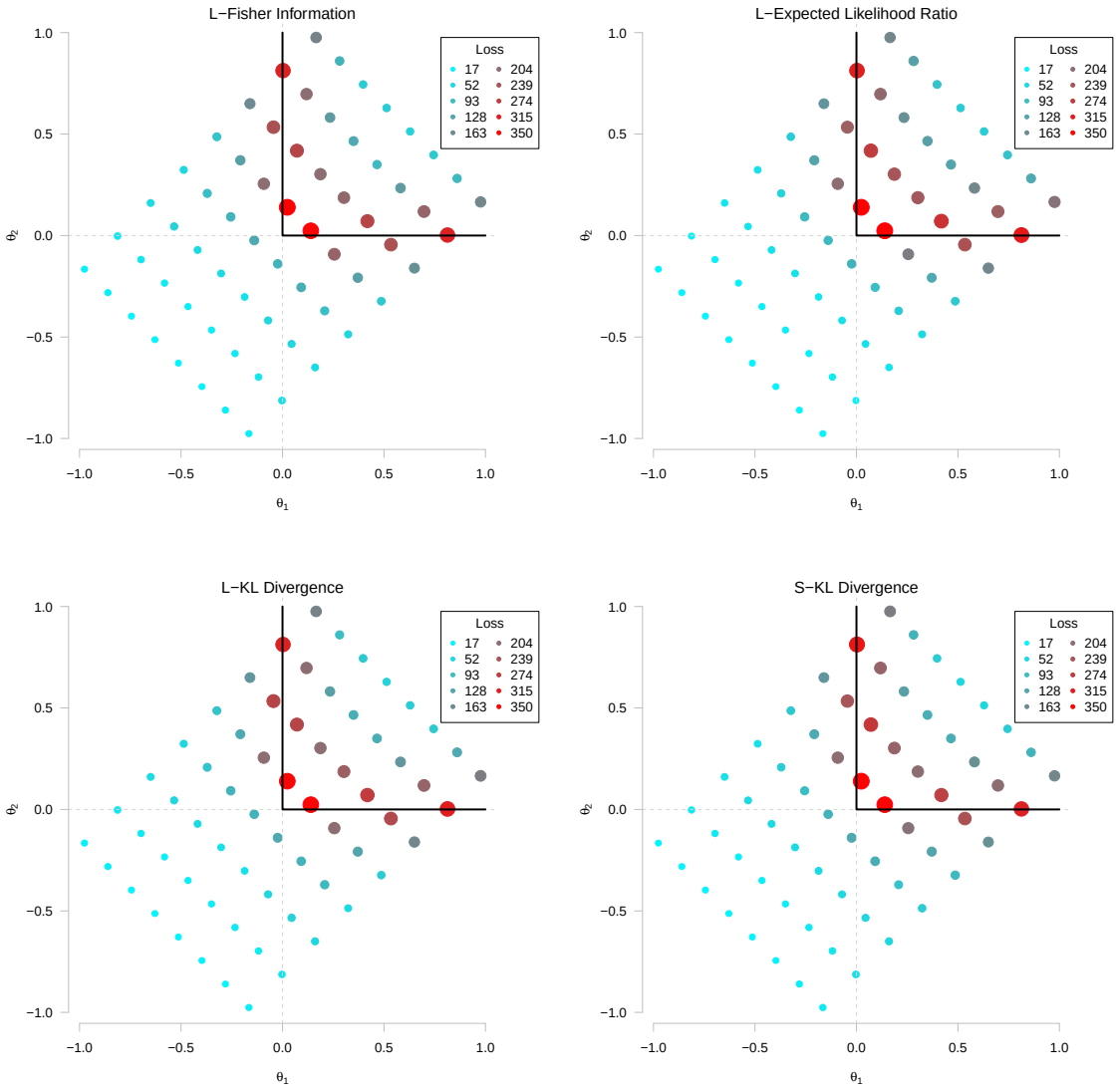

Figure D.37: Scatterplots of the conditional average loss for various vectors of true ability when using the non-compensatory classification bound function and the BCR stopping rule with $\alpha = .10$. Different panels represent different item selection algorithms. Bubbles are color-coded and sized according to loss. See the right-most panel of Figure D.1 for more information.