The Expected Likelihood Ratio in Computerized Classification Testing

Steven W. Nydick

Pearson VUE

-

Author Note

Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME) in Washington D.C., April 7–11, 2016.

Abstract

The Sequential Probability Ratio Test (SPRT) is commonly used as a stopping rule for computerized adaptive mastery tests (CMT). Many researchers claim that items should be selected for CMT by maximizing Fisher information at the bound separating two categories. In Nydick (2014), we demonstrated that the optimal location for item selection in adaptive CMT depends on both the classification bound as well as the latent trait underlying item responses. The Expected Likelihood Ratio (ELR) method of item selection, which optimizes the SPRT-based likelihood ratio test statistic with respect to a person's current ability estimate, improved over standard cut-point based algorithms in terms of average test length with little change in classification accuracy (for the limited conditions tested in that paper). The current study extends Nydick (2014) to systematically compare the ELR with alternative item selection algorithms using a wide variety of item parameter banks as well as cut-points. Results from this simulation study support claims from Nydick (2014): Item selection algorithms that considered both the cut-point as well as the ability estimate performed better, in terms of average test length and without a loss in classification accuracy, than algorithms that considered either piece of information alone.

The Expected Likelihood Ratio in Computerized Classification Testing

Introduction

Computerized classification tests (CCT; e.g., Eggen, 1999) are computerized adaptive tests (CAT; e.g., Weiss, 1982) that aim to accurately and efficiently assign examinees into categories. Most research on CCT has focused on stopping rules. Commonly used stopping rule algorithms include the confidence interval method (CI; e.g., Kingsbury & Weiss, and also referred to as Sequential Bayes, e.g., Spray & Reckase, 1996) and the sequential probability ratio test (SPRT; e.g., Eggen, 1999), although alternative methods have been proposed to alleviate CI or SPRT inefficiencies (e.g., Finkelman, 2003; Huebner & Fina, 2015; Sie, Finkelman, Bartroff, & Thompson, 2014; Thompson, 2009).

The SPRT is a straightforward stopping rule taken from statistical decision theory (Wald, 1947). Assume that we are conducting a sequence of identical experiments with simple, point hypotheses about a parameter of interest. Moreover, assume that we have fixed critical values, A and B, such that $0 < A < B < \infty$. After each step of the experiment, we need to make a decision about whether either of the point hypotheses are true or whether to perform another experiment and collect another observation. According to the Wald-Wolfowitz theorem (Wald & Wolfowitz, 1948), the optimal test statistic (i.e., the test statistic the results in the least number of experiments/the smallest sample size under either hypothesis and with the same Type I and Type II error rates) is the simple likelihood ratio.

To define the SPRT in CCT, one only needs to define the null and alternative hypotheses and specify critical regions. Let θ represent the latent ability variable modeled by a unidimensional IRT model, and denote an instance of θ as θ_i. (θ can be thought of as ability in general, whereas θ_i is a particular person's ability.) Moreover, define a cut-point on an exam as θ_0. Then any stopping rule would seek to determine, after each question, whether $\theta_i > \theta_0$, $\theta_i < \theta_0$, or whether not enough information has been collected from person i to make a decision. Because the SPRT requires point hypotheses (instead of the general "somewhere above the cut-point"), let each hypothesis be determined by a point slightly away from the cut-point. In other words, we can specify point hypotheses as

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

where δ is a small constant.

After an examinee responds to a test item, the CAT algorithm compares the likelihood of the sequence of responses for this examinee (to the current point in the test) given H_0 versus H_1. Assuming responses to a set of test items are conditionally independent, then the log-likelihood function for θ given response pattern $\mathbf{y}_i$ is

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

where y_{ij} is the (0, 1) response of examinee i to item j and $p_j(\theta)$ is probability of $y_{ij} = 1$ given a value of θ (and defined by the IRT model). If testing the point hypotheses defined above, then the log-likelihood ratio of an examinee having ability θ_u relative to θ_l is

$$\log\Big[\mathrm{LR}(\theta_u, \theta_l|\mathbf{y}_i)\Big] = \log\left[\frac{L(\theta_u|\mathbf{y}_i)}{L(\theta_l|\mathbf{y}_i)}\right] = \log\Big[L(\theta_u|\mathbf{y}_i)\Big] - \log\Big[L(\theta_l|\mathbf{y}_i)\Big]. \tag{2}$$

When Equation (2) is greater than $\log(B)$ or less than $\log(A)$, then the test is ended and the appropriate hypothesis chosen. Otherwise, another item is given until maximum test length. Typically $A = [\beta/(1-\alpha)]$ and $B = [(1-\beta)/\alpha]$, which results in Type I and Type II error rates approximately α and β, respectively (e.g., Wald, 1945).

Regardless of whether the CI or SPRT methods are used to terminate a classification test, most researchers assume that items should be selected with maximum information at the bound separating categories (e.g., Eggen, 1999; Finkelman, 2008; Huebner, 2012). In Nydick (2014), we demonstrated that "selecting items to maximize information at the classification bound is only a coarse approximation of the most efficient item selection algorithm" with respect to the SPRT (p. 206). The next section summarizes the main results from that paper. Based on these results, we explore a simulation study designed to systematically test whether one can improve on the classic maximize information at the cut-point item selection algorithm in SPRT-based classification tests.

Item Selection in Mastery Tests

Many researchers select items in CCT algorithms by maximizing information at the cut-point separating categories. The main reasons for this particular recommendation is intuitive and obvious. When maximizing information at the cut-point, the difficulty of items will be close to the cut-point. Therefore, if a candidate answers a question correctly, then that candidate is most likely in the region above the cut-point, and if the candidate answers a question incorrectly, then that candidate is most likely in the region below the cut-point. Unfortunately, at least with respect to the SPRT, this reasoning is not entirely accurate. In Nydick (2014), we demonstrated that selecting items by considering both the classification bound as well as the current ability estimate can improve on the standard item selection algorithm. This section explains how an examinee's estimated ability can improve item selection in SPRT-based classification tests.

For simplicity, assume that we only have two categories separated by a cut-point, θ_0, and assume that all items conform to a unidimensional, binary three-parameter logistic (3PL) model. Moreover, let θ_i be examinee i's latent ability, and assume that all item responses are conditionally independent given θ_i. Then we can represent the probability of examinee i correctly responding to item j with the following item response function (IRF):

$$p_j(\theta_i) = \Pr(Y_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + \exp[-Da_j(\theta_i - b_j)]}, \tag{3}$$

where b_j determines the point of maximal slope, a_j is proportional to the slope at $\theta_i = b_j$, c_j represents the lower asymptote, and D is a scaling constant. Because D is a scaling constant that does not affect model fit, we remove D from subsequent equations and define a_j as being implicitly multiplied by D. Common simplifications of Equation (3) include the two-parameter logistic (2PL) model, where $c_j = 0$ for all items, and the one-parameter logistic (1PL) model, where $a_j = 1$ for all items.

Now assume that examinee i has just taken an item conforming to Equation (3), and we want to determine whether that examinee should be classified in the upper category (e.g., $\theta_i = \theta_0 + \delta$), the examinee should be classified in the lower category (e.g., $\theta_i = \theta_0 - \delta$), or the examinee should be given another item. Then according to the SPRT, we will form the log-likelihood ratio (comparing $\theta_0 + \delta$ to $\theta_0 - \delta$) of all items to this point in the test and determine whether that ratio is outside of our critical bounds. Based on Equation (1), the log-likelihood ratio is summative under our assumptions, so we can determine the incremental evidence of question j, which, of course, depends on the item response,

$$\begin{aligned}\log\Big[\mathrm{LR}(\theta_0+\delta, \theta_0-\delta|y_{ij})\Big] &= \log\left[\frac{L(\theta_0+\delta|y_{ij})}{L(\theta_0-\delta|y_{ij})}\right] \\ &= y_{ij}\log\left[\frac{p_j(\theta_0+\delta)}{p_j(\theta_0-\delta)}\right] + (1-y_{ij})\log\left[\frac{1-p_j(\theta_0+\delta)}{1-p_j(\theta_0-\delta)}\right]. \end{aligned} \tag{4}$$

In Nydick (2014), we demonstrated that Equation (4) behaves differently for correct versus incorrect answers, as long as $c_j > 0$. One can see the difference by assuming a correct or incorrect response and then determining the classification bound that provides optimal evidence for classification. If $y_{ij} = 0$ (person i answers item j incorrectly), then Equation (4) is monotonically decreasing with respect to θ_0, so that higher cut-points result in more incremental evidence for person i being in the lower category. This result aligns with intuition: if an examinee is classified into the lower category given $\theta_0 = 4$ and an incorrect response, that examinee should also be classified into the lower category given $\theta_0 = 10$ or $\theta_0 = 20$ or $\theta_0 = 256$. Unfortunately, if $y_{ij} = 1$ (person i answers item j correctly), then the behavior of Equation (4) does not necessarily accord with intuition.

Assuming $y_{ij} = 1$, Equation (4) becomes

$$\log\left[\mathrm{LR}(\theta_0 + \delta, \theta_0 - \delta | y_{ij} = 1)\right] = \log\left[\frac{p_j(\theta_0 + \delta)}{p_j(\theta_0 - \delta)}\right] = \log[p_j(\theta_0 + \delta)] - \log[p_j(\theta_0 - \delta)], \quad (5)$$

and the maximum of Equation (5) with respect to θ_0 is (see Equation 11 in Nydick, 2014)

$$\hat{\theta}_0 = \frac{\log(c_j)}{2a_j} + b_j. \quad (6)$$

If $c_j = 0$, then the optimal classification bound does indeed approach negative infinity (which is analogous to the incorrect response case). However, if $c_j > 0$, then item j will give optimal incremental evidence for classification if θ_0 is just a bit below b_j. For instance, let $c_j = .25$, so that item j represents the ideal, four choice multiple choice question. Then when $a_j = 1$, the optimal classification bound is approximately .69 below an item's difficulty. When $a_j = 2$, then the optimal classification bound is only .35 below an item's difficulty. Therefore, if an examinee answers a question incorrectly, classification evidence will always be stronger for higher classification bounds. However, if an examinee answers a question correctly, then the classification evidence depends highly on the parameters of the item. Although this result sounds counterintuitive, we explained (in Nydick, 2014) that it is entirely a consequence of the SPRT test statistic being a simple likelihood ratio. If θ_0 is far below b_j, then the likelihood ratio is essentially 0 because candidates both at $\theta_0 + \delta$ and $\theta_0 - \delta$ will essentially answer question j at guessing, so that $\frac{c_j + \epsilon_1 + \epsilon_2}{c_j + \epsilon_1} \approx 1$, where ϵ_1 and ϵ_2 are small constants indicating the incremental probability of examinees at ability $\theta_0 - \delta$ (relative to $-\infty$) and $\theta_0 + \delta$ (relative to $\theta_0 - \delta$) answering the item correctly.

One problem with Equation (6) is the assumption of a correct response. Typically, candidate responses are not known prior to actually administering an item. However, one can also determine the optimal item to administer given an expected response to an item,

$$\mathbb{E}_{Y_{ij}|\theta_i}\left[\log\left[\text{LR}(\theta_0+\delta,\theta_0-\delta|Y_{ij})\right]\right] = p_j(\theta_i)\log\left[\frac{p_j(\theta_0+\delta)}{p_j(\theta_0-\delta)}\right] + [1-p_j(\theta_i)]\log\left[\frac{1-p_j(\theta_0+\delta)}{1-p_j(\theta_0-\delta)}\right]. \tag{7}$$

Equation (7) does not depend on the actual response to an item but considers how examinees might respond across many administrations given items of similar difficulty. Note that Equation (7), or the Expected log-Likelihood Ratio (ELR), is equivalent to the expectation of the SPRT test statistic. An item that optimizes Equation (7) would be expected to contribute maximum evidence for classification. In Nydick (2014), we demonstrated that under a few simplifying assumptions (including $c_j = 0$, fixed and constant a_j, and δ approaching 0), the optimal item would have a difficulty of

$$\lim_{\delta\to 0^+} \hat{b} = \frac{\theta_0+\theta_i}{2}. \tag{8}$$

Therefore, the item providing the best evidence for classification would not have difficulty close to the classification bound but approximately halfway between the classification bound and a person's actual ability.

To demonstrate the superiority of Equation (8), we constructed a fairly simple simulation study using three item selection algorithms: maximize Fisher information at the classification bound, maximize Fisher information at the current ability estimate, and maximize Fisher information halfway between the classification bound and the current ability estimate. Item parameters were selected from a bank of $J = 750$ items with b_j uniformly distributed between -4 and 4, a_j distributed in the manner described in the next section, and $c_j = 0$. Note that this simulation study represents a select few conditions of the broader simulation study described below. For a better understanding of the manipulated conditions or the distributions therein, see this paper's method section. Note that $N = 10,000$ simulees were generated such that $\theta \sim N(0,1)$.

Figure 1 indicates the average test length across $N = 10,000$ simulees given seven

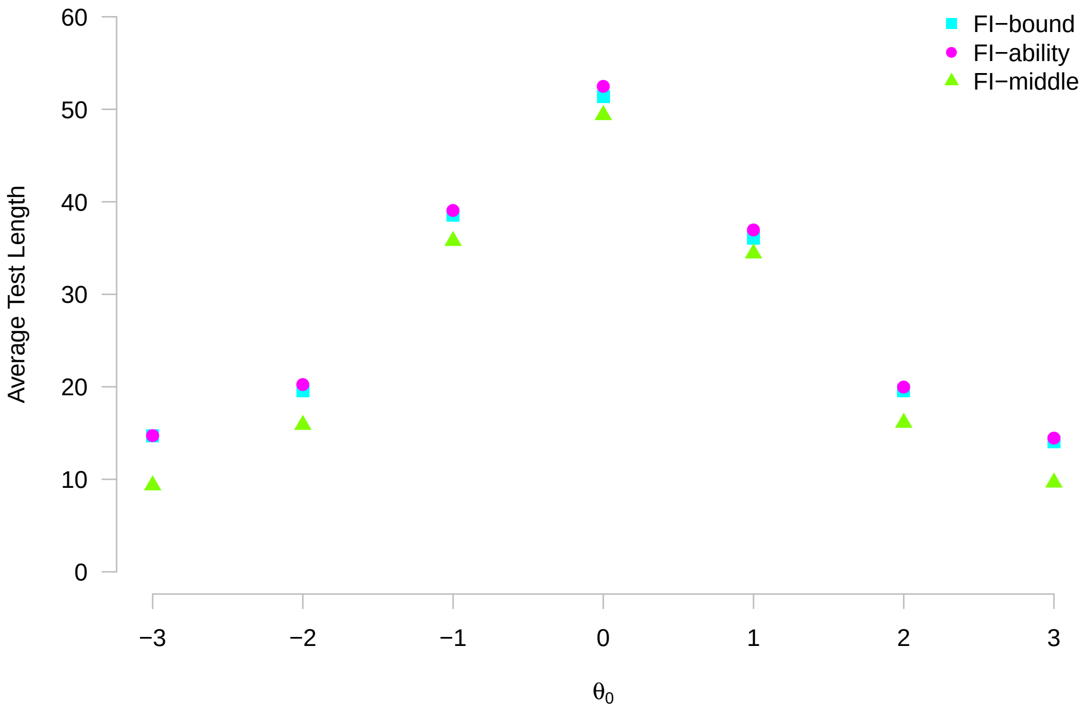

Figure 1. Average test length using an SPRT stopping rule with item selection algorithms and classification bounds with items generated according to the 2PL IRT model. For each combination of item selection algorithm and classification bound, test length was averaged across $N = 10,000$ simulees generated from a $N(0,1)$ distribution with a minimum test length of $j_{\text{min}} = 5$ and a maximum test length of $j_{\text{max}} = 200$. For the SPRT stopping rule, $\delta = .1$ and $\alpha = \beta = .05$.

different classification bounds. The classification bounds were applied separately, such that individual simulations were run with each of the classification bounds, and the purpose of the SPRT algorithm was to simply decide if a candidate did or did not exceed a cut-point. Notice that FI-bound (i.e., maximizing information at the classification bound) and FI-ability (i.e., maximizing information at the current ability estimate) performed similarly with respect to average test length. Given a cut-point toward the middle of the distribution, FI-bound resulted in an average test length slightly shorter than FI-ability: the blue square is slightly below the purple circle for cut-points of $\theta \in \{-1, 0, 1\}$. However, the differences between these two conditions are fairly minor. A major reason for their similar performance is due to the missing lower asymptote. If $c_j > 0$, then FI-bound should perform much better than FI-ability, in terms of average test length, for extremely negative θ_0. Yet FI-middle (i.e., maximizing information halfway between the classification bound and the current ability estimate) resulted in shorter tests than the other two conditions. The difference between test length of FI-middle versus FI-bound or FI-ability ranged anywhere from 5 items if $\theta_0 = -3$ or $\theta_0 = 3$ to 2 items if $\theta_0 = 0$. This simple combination of two standard item selection algorithms resulted in shorter average tests across all cut-points on tests without lower asymptotes, just as Equation (8) predicted.

Figure 2 indicates the percent misclassified across the conditions described above. Note that the misclassification across given different item selection algorithms are comparable, and FI-middle performed better than (given $\theta_0 = 0$) or at almost identically to (for the other six classification bounds) the standard item selection algorithms. This simple simulation reinforces the derivations from Nydick (2014), in that selecting items at $b \approx \frac{\theta_0 + \hat{\theta}_i}{2}$ resulted in a larger expected change in the SPRT test statistic (and, thus, a shorter test) than selecting items at $b \approx \theta_0$ or $b \approx \hat{\theta}_i$.

Unfortunately, FI-middle is based on Equation (8), which is an asymptotic result. Therefore, this particular selection algorithm is only guaranteed to lead to relatively efficient tests given models with $c_j = 0$ and an SPRT routine with small δ. In Nydick (2014), we demonstrated how the optimal item changes for candidates with ability above or below the classification bound for various values of c_j and δ. This demonstration led to the proposal of an item selection algorithm for SPRT-based computerized mastery tests that would retain optimality properties without requiring stringent assumptions, such as a particular model or indifference region. Rather than maximize Fisher information halfway between ability and the classification bound, one could simply optimize Equation (7), the ELR. Because a person's ability would never be known, one could replace θ_i in Equation (7) with $\hat{\theta}_i$. This ELR algorithm (Nydick, 2014, p. 215) would then choose subsequent

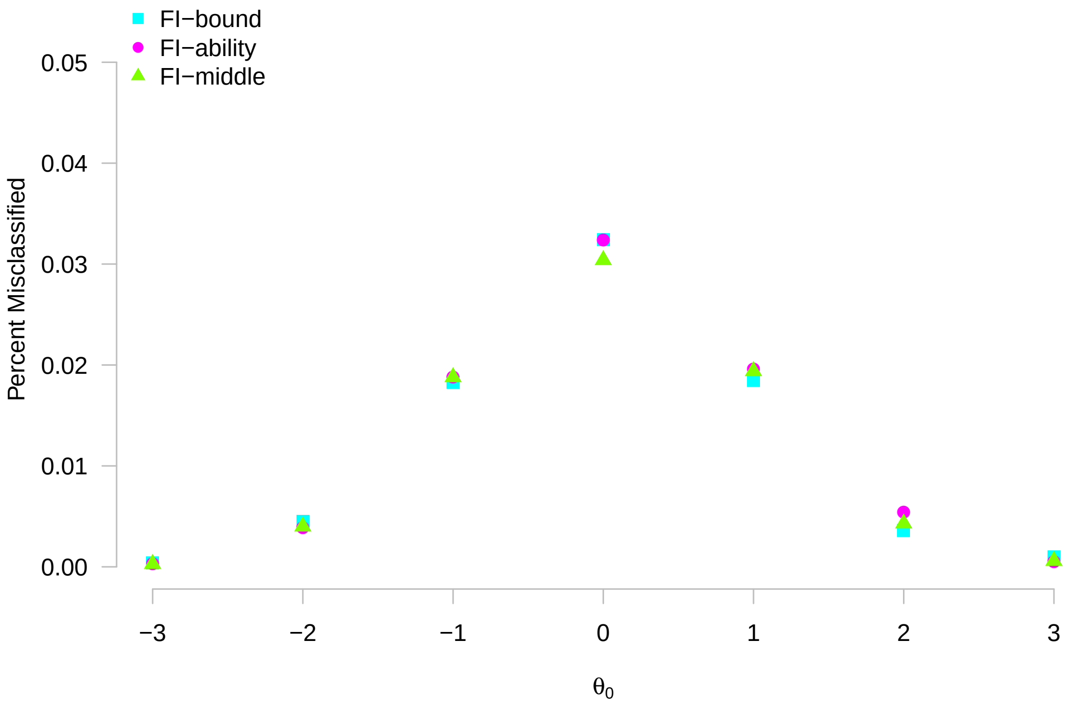

Figure 2. Percent misclassified using an SPRT stopping rule with item selection algorithms and classification bounds with items generated according to the 2PL IRT model. For each combination of item selection algorithm and classification bound, misclassification rate was averaged across $N = 10,000$ simulees generated from a $N(0, 1)$ distribution with a minimum test length of $j_{\text{min}} = 5$ and a maximum test length of $j_{\text{max}} = 200$. For the SPRT stopping rule, $\delta = .1$ and $\alpha = \beta = .05$.

items by maximizing Equation (7) if $\hat{\theta}_i > \theta_0$ and minimizing Equation (7) if $\hat{\theta}_i < \theta_0$.

In Nydick (2014), we conducted several simulations to demonstrate the plausibility of the ELR in choosing items for a SPRT-based classification test. However, the simulations from that paper only demonstrated the possibility of improving on the standard cut-point-based algorithms. They were not constructed to comprehensively test the superiority of the ELR. Moreover, alternative item selection algorithms might be constructed that also consider estimated ability as well as the cut-point in making decisions. The purpose of this paper is to extend Nydick (2014) by comprehensively testing several item selection algorithms, including the ELR, in a simulation study. In the remainder of this paper, we outline several novel algorithms that consider both the cut-point as well as the ability estimate in choosing subsequent items. We then describe a simulation study that compares the average test length and classification accuracy when using these algorithms and explain implications of this simulation for developing item selection algorithms for computerized classification tests.

Simulation Study

The current section describes a simulation study designed to compare the efficiency and accuracy of the SPRT in classifying simulees when using a variety of item selection algorithms, including the ELR. For all simulation conditions, the minimum exam length was $j_{\text{min}} = 5$, and the maximum exam length was $j_{\text{max}} = 200$. Moreover, the simple SPRT was used to classify simulees into one of two categories with $\delta = .1$ and $\alpha = \beta = .05$ (see Equations 2 and 4). Finally, $\hat{\theta}$ (needed for some of the algorithms) was restricted to be between $\hat{\theta} \in \{-6, 6\}$. ($\hat{\theta}$ was set to the maximum or minimum of this range for non-mixed response patterns.)

Person Parameter Generation

All simulation conditions contained the same $N = 10,000$ simulees, such that $\theta \sim N(0, 1)$. The same simulees were used for all conditions to reduce some of the random noise that can be generated given different groups of simulees. Response vectors were always randomly and separately generated for each combination of conditions, even if the simulees saw the same item bank.

Item Parameter Generation

All items were generated in accordance with the 3PL IRT model. Eighteen different item banks were generated and administered to simulees. Conditions varied included the distribution of item difficulties, the distribution of item lower asymptotes, and the number of items.

Item difficulties were randomly distributed based on three specifications: “flat”, “moderate”, or “peaked”. If the item difficulty distribution was “flat”, then b-parameters were generated from a uniform distribution with a minimum of -4 and a maximum of 4. This distribution should result in a reasonable set of items for most simulees and is in line with standard CAT recommendations (e.g., Weiss, 1982, p. 478). If the item difficulty distribution was “moderate” or “peaked”, then b-parameters were generated from a normal distribution with a mean of 0 and a standard deviation of 1.500 (if “moderate”) or 0.707 (if “peaked”). The latter two distributions should result in fewer possible items for candidates with extremely low or high values of θ.

Item lower asymptotes were either randomly distributed, set to 0.25, or set to 0.00. Under the “random” condition, c-parameters were generated from a beta distribution with $c \sim \text{Beta}(\alpha = 19.8, \beta = 79.2)$. This distribution would result in c-parameters with a mean of approximately .20, a standard deviation of approximately .04, and almost all values contained between .10 and .30.

Item discriminations were always randomly generated from the log-normal distribution with $a \sim \log N(\mu_{\text{log}} = 0.38, \sigma_{\text{log}} = 0.25)$ (which is similar to the distribution used in Nydick, 2014).

For each combination of parameter conditions, two item banks were generated, one with $J = 750$ items and one with $J = 1,500$ items. These item banks remained constant for all simulations containing them. In other words, given a particular combination of item bank conditions (e.g., “flat” b, “random” c, and $J = 750$), one item bank was generated, and that bank was used for all simulation conditions containing that set of item bank conditions. As was the case in generating person parameters, keeping the item banks constant for all sets of relevant conditions reduces nuisance variance and provides a better base for comparing simulation results.

Item Selection Algorithms

The simulation was designed to test various item selection algorithms, and we incorporated eight algorithms in the simulation design, including three standard algorithms (FI-bound, FI-ability, and KL-bound), two modified algorithms (FI-middle and KL-estimated), one novel algorithm (ELR), and two hybrid algorithms (FI-ability-hybrid and KL-estimated-hybrid).

FI-bound, FI-ability, and FI-middle selected items to maximize Fisher information (which is the expected, negative, second derivative of the log-likelihood function and relates to the asymptotic variance of the maximum likelihood estimator; see Efron & Hinkley, 1978). As described earlier, FI-bound chooses items to maximize information at the classification bound, FI-ability chooses items to maximize information at the ability estimate, and FI-middle chooses items to maximize information halfway between the classification bound and the ability estimate.

In addition to Fisher information-based algorithms, several alternative item selection algorithms have been proposed for adaptive classification tests (e.g., Eggen, 1999; Lin & Spray, 2000; Weissman, 2007). One commonly-used alternative to Fisher information is based on Kullback-Leibler (KL) divergence (Chang & Ying, 1996). KL-divergence quantifies the expected loss when using an approximate model in lieu of the correct model by means of a log-likelihood ratio. Most KL-divergence-based algorithms in CCT choose the upper hypothesis ($\theta_0 + \delta$) as the true distribution (e.g., Eggen, 1999), which results in the following maximization function,

$$\begin{aligned}\mathrm{KL}_j(\theta_0+\delta||\theta_0-\delta) &= \mathbb{E}_{Y_{ij}|\theta_0+\delta}\Big[\log\Big[\mathrm{LR}(\theta_0+\delta,\theta_0-\delta|Y_{ij})\Big]\Big] \\ &= p_j(\theta_0+\delta)\log\left[\frac{p_j(\theta_0+\delta)}{p_j(\theta_0-\delta)}\right] + [1-p_j(\theta_0+\delta)]\log\left[\frac{1-p_j(\theta_0+\delta)}{1-p_j(\theta_0-\delta)}\right]. \qquad (9)\end{aligned}$$

Note that Equation (9) is similar to Equation (7) with θ_i replaced by $\theta_0 + \delta$. A problem with the KL-bound algorithm is that $\text{KL}_j(\theta_u||\theta_l)$ assumes that all candidates are in the upper region. One could construct an alternative algorithm by flipping $\theta_0 + \delta$ and $\theta_0 - \delta$ in Equation (9) if an examinee's ability estimate is below the classification bound. The KL-bound algorithm chooses items to maximize Equation (9), whereas the KL-estimated algorithm chooses items to maximize $\text{KL}_j(\theta_u||\theta_l)$ if $\hat{\theta}_i > \theta_0$ and $\text{KL}_j(\theta_l||\theta_u)$ if $\hat{\theta}_i < \theta_0$.

The hybrid algorithms consider that $\hat{\theta}_i$ is an imprecise estimate of θ_i. Therefore, choosing items based on the ELR will be non-optimal as long as $|\hat{\theta}_i - \theta_i|$ remains large. The FI-ability-hybrid algorithm and KL-estimated-hybrid algorithm choose items based on FI-ability or KL-estimated until the standard error of $\hat{\theta}_i$ is small enough, and then switch to the ELR algorithm for the remainder of the CAT. In all cases, we applied a standard error threshold (before switching to the ELR algorithm) of 0.5. A standard error of 0.5 is large enough for the ELR to be applied relatively early in a test but small enough for the coarse interval of θ_i to be mostly known.

Conditions

Ultimately, we included 18 item banks and 8 item selection algorithms. For each of the $18 \times 8 = 144$ bank by item selection algorithm conditions, we simulated computerized classification tests for each of the $N = 10,000$ simulees at each of $\theta_0 \in \{-3, -2, -1, 0, 1, 2, 3\}$ classification bounds, which resulted in a $144 \times 7 = 1008$ total bank × item selection algorithm × classification bound conditions. Given a condition, we generated adaptive classification tests (using the global parameters) across the $N = 10,000$ simulees and determined the average test length and classification accuracy for the simulee group. The next section describes results from this simulation study.

Simulation Results

This section will be divided into two parts: presenting results aggregated across all simulees within a particular condition, and presenting results for simulees with similar values of θ. In all cases, only a select few results will be presented, entirely in graphical form. The full set of results (including graphs and tables) will be uploaded to www.tc.umn.edu/~nydic001. Moreover, we calculated loss values (e.g., Finkelman, 2010) as

$$\text{Loss} = P \times I_W + J, \tag{10}$$

where I_w is an indicator function for an incorrect classification, J is the test length, and P is a penalty applied to an incorrect classification. In most cases, and with $P \in \{100, 500\}$, loss did not indicate a different rank order than simply the average test length. Therefore, we will use test length as a proxy for loss for the remainder of the document.

Aggregated Across All Simulees

Figure 3 presents the average test length deviance for each item selection algorithm and classification bound from the item bank where $J = 750$, $c = .25$, and b uniformly distributed between -4 and 4. The average test length deviance for a condition is simply the difference between the average number of items administered given an item selection algorithm and a classification bound and the median of the average number of items administered across all item selection algorithms for that classification bound. The reason that the deviance was chosen to describe average test length (rather than the straightforward average test length) is that the average test length magnitude depends on the classification bound. The average test length would be much larger for classification bounds closer to $\theta = 0$ than $\theta = 3$ simply because more simulees are closer to $\theta = 0$, and a variable length classification test takes longer to make a decision if a simulee/examinee is close to the classification bound. The deviance statistic removed the difference in location across all θ_0 and made the different conditions more comparable. The reason that the median was chosen for the deviance rather than the mean is simply because several item selection algorithms were expected to result in very large values for the average test length, and the median is less susceptible to outliers.

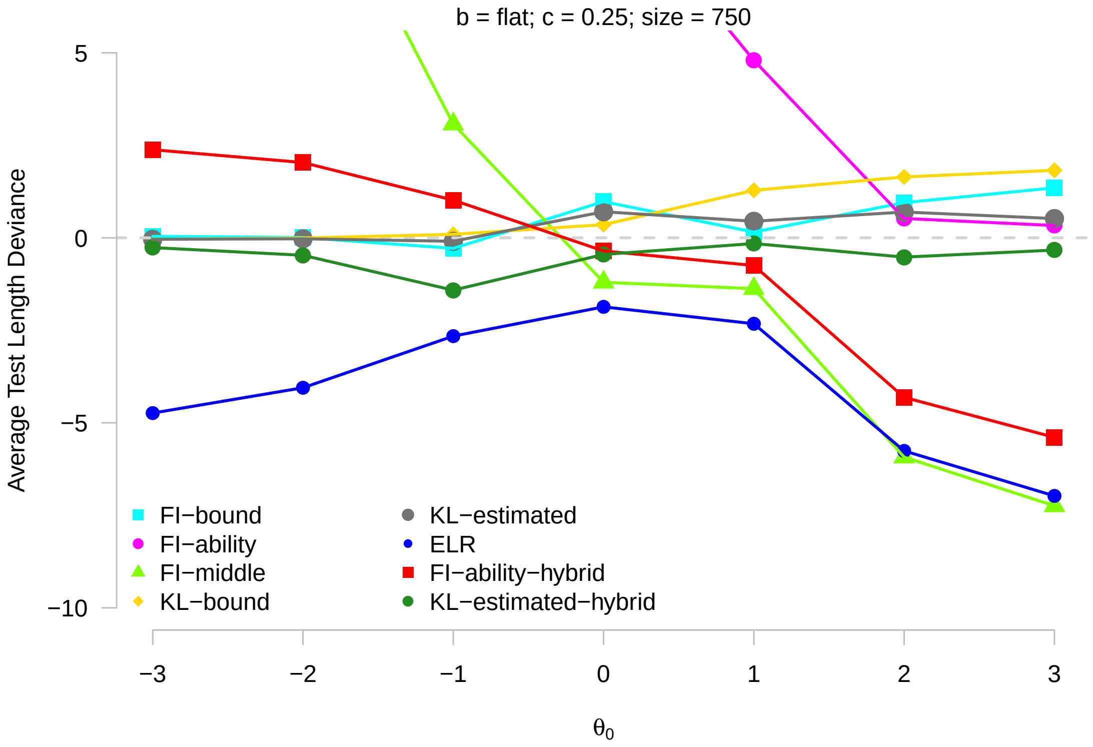

Figure 3. Average test length deviance using an SPRT stopping rule with item selection algorithms and classification bounds with items generated according to the 3PL IRT model with $J = 750$ items generated from a bank such that $c = .25$ and b uniformly distributed between -4 and 4.

The results displayed in Figure 3 are almost archetypally similar to what one would predict. First, the three algorithms that depend on maximizing Fisher information based on the ability estimate (FI-ability, FI-middle, and FI-ability-hybrid) result in much longer tests for very negative classification bounds, which is expected for item banks with positive lower asymptotes. FI-ability and FI-middle perform so much worse than the other conditions if $\theta_0 << 3$ that their average test length deviance is outside the range of the y-axis on the graph. However, FI-middle performs just as well as the ELR item selection algorithm if $\theta_0 = 2$ or $\theta_0 = 3$. For these classification bounds, most simulees are below the classification bound, so the effect of the lower asymptote on the SPRT statistic is much less. In fact, all three item selection algorithms perform much better for $\theta_0 = 3$ than for any of the other conditions.

Next, notice that KL-estimated (the grey circles) performs similarly to KL-bound (the yellow triangles) for $\theta_0 <= 0$ but results in shorter tests than KL-bound for $\theta_0 > 0$. The difference between the KL conditions is not dramatic. However, considering the location of θ_i relative to θ_0 does result in added efficiency, especially when most simulees have true ability below the classification bound. Finally, look at the ELR condition and the hybrid conditions. The ELR results in the shortest tests for most conditions, and FI-ability-hybrid results in tests nearly as short for $\theta_0 \geq 0$.

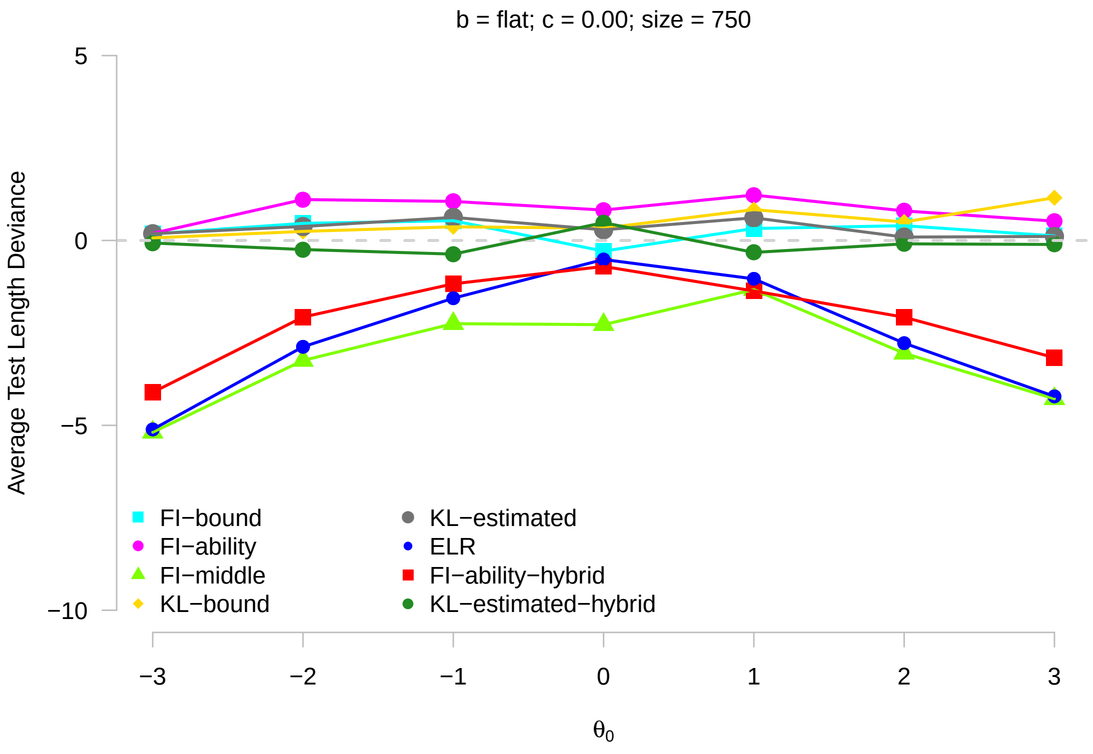

Figure 4. Average test length deviance using an SPRT stopping rule with item selection algorithms and classification bounds with items generated according to the 2PL IRT model with $J = 750$ items generated from a bank such that $c = 0$ and b uniformly distributed between -4 and 4.

Figure 4 displays the average test length deviances for the item bank similar to that shown in Figure 3 (with $J = 750$ and b uniformly distributed between -4 and 4) but with $c = 0$. Notice that the item selection algorithms perform much more similarly if $c = 0$ than if $c > 0$. Because $c = 0$ in Figure 4, FI-ability is not penalized for high ability simulees relative to the classification bound. Therefore, none of the lines have a dramatic leftward increase in test length deviance, unlike the patterns shown in Figure 3. Interestingly, although ELR performs better than most of the item selection algorithms (including all of the standard algorithms) in terms of test length deviance, FI-middle performs almost uniformly better than the ELR. Notice that the bright green triangles are always at least as low as the dark blue circles.

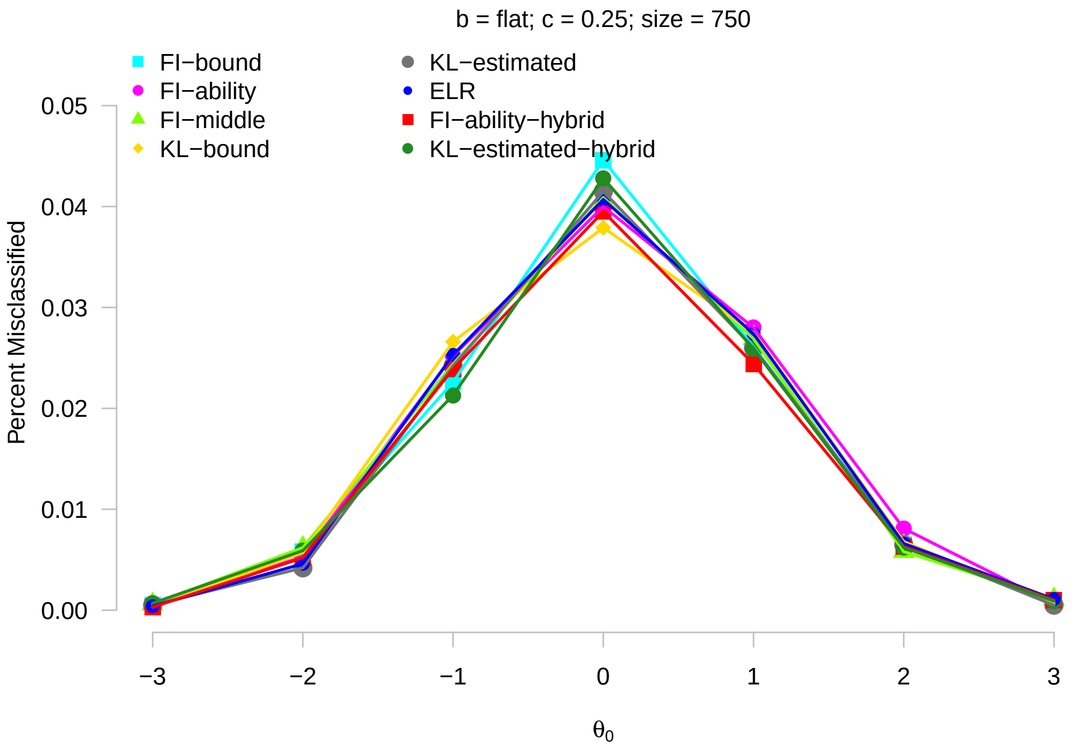

Figure 5. Percent misclassified using an SPRT stopping rule with item selection algorithms and classification bounds with items generated according to the 3PL IRT model with $J = 750$ items generated from a bank such that $c = .25$ and b uniformly distributed between -4 and 4.

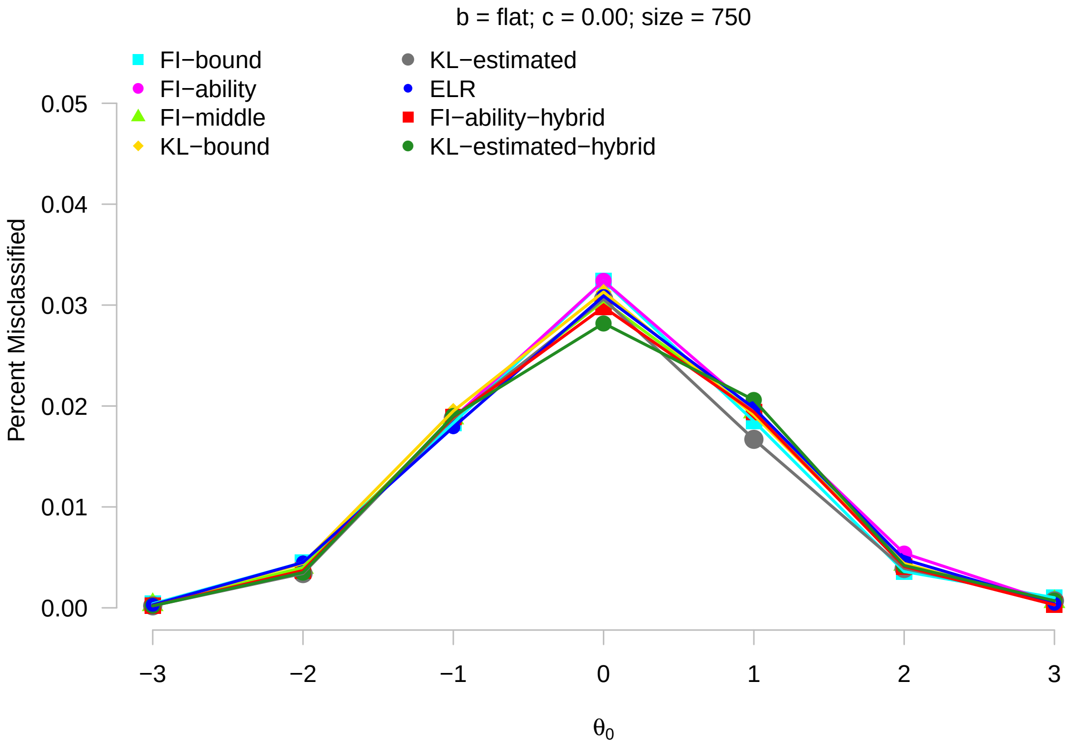

Figure 6. Percent misclassified using an SPRT stopping rule with item selection algorithms and classification bounds with items generated according to the 2PL IRT model with $J = 750$ items generated from a bank such that $c = 0$ and b uniformly distributed between -4 and 4.

Figures 5 and 6 display the misclassification rate for the conditions described by Figures 3 and 4. Notice that the misclassification rates across all item selection algorithms (for a given classification bound) are very similar: one finds little consistency in the relative order of the item selection algorithms in terms of misclassification rate across the different classification bounds. The item selection algorithms should not result in drastically different misclassification rates, as they only determine the optimal item to administer and not whether the test should end. The misclassification rate should depend almost entirely on the stopping rule. Otherwise, the Type I and Type II errors would depend on the type of evidence collected rather than simply the amount of evidence.

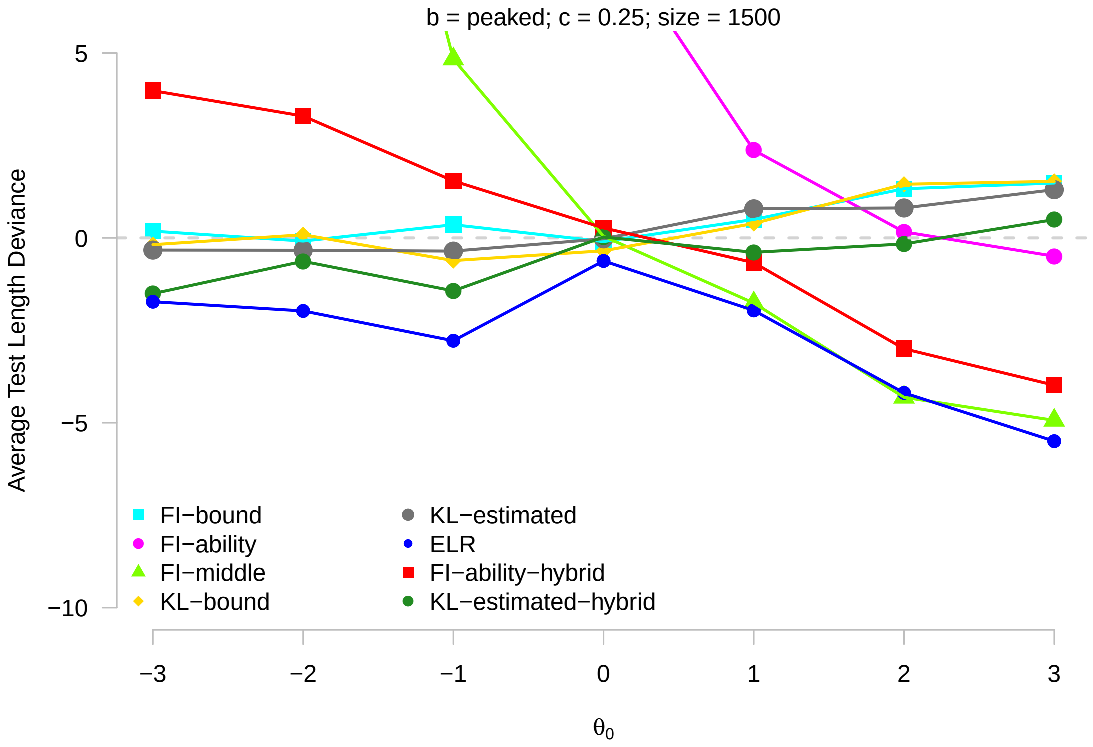

Figure 7. Average test length deviance using an SPRT stopping rule with item selection algorithms and classification bounds with items generated according to the 3PL IRT model with $J = 1,500$ items generated from a bank such that $c = .25$ and b normally distributed with a mean of 0 and a standard deviation of 0.707.

Figures 7 and 8 display average test length deviances for all of the item selection algorithms at each of the classification bounds given an item bank with more items ($J = 1,500$) but a much narrower difficulty range ($b \sim N(\mu_b = 0, \sigma_b = 0.707)$). The difference between Figures 7 and 8 is that in Figure 7, $c = .25$ for all items, and in Figure 8, $c = 0$. Notice that Figures 7 and 8 are similar to Figures 3 and 4 with a few exceptions. For instance, given a peaked item bank, most of the item selection algorithms result in tests lengths closer to the median test length than with a wider range of difficulties. This result makes intuitive sense. If b is generated from a narrow range, fewer items will have difficulties close to extreme values of θ regardless of whether selecting items to maximize

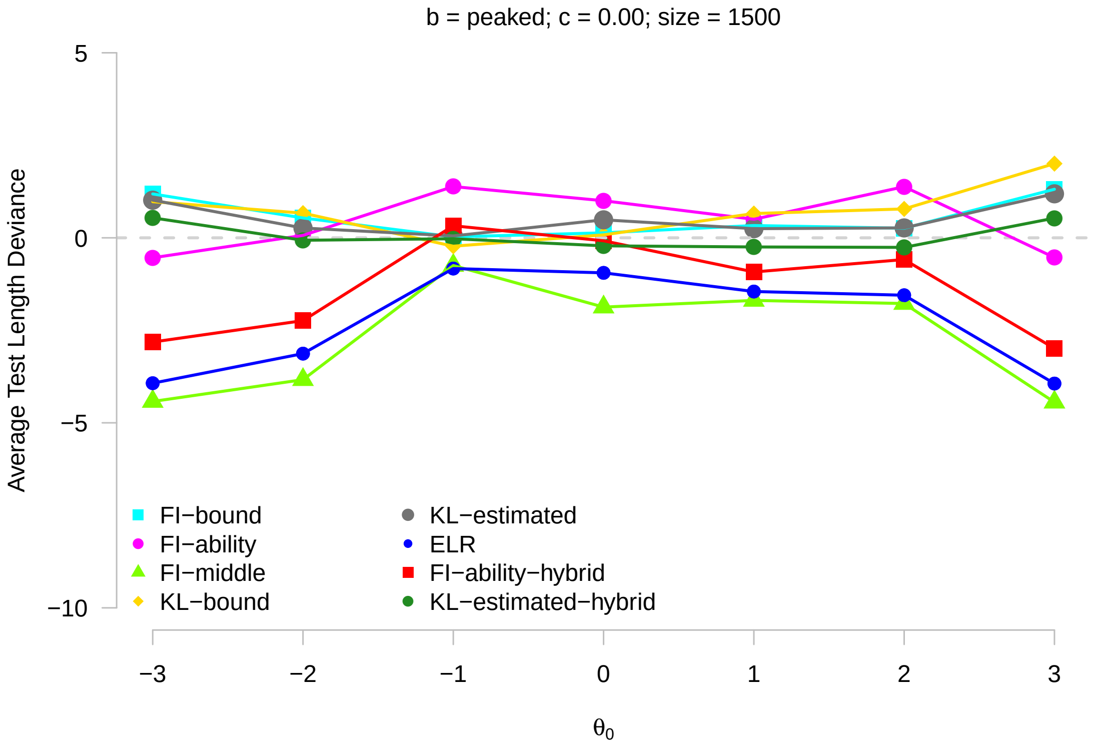

Figure 8. Average test length deviance using an SPRT stopping rule with item selection algorithms and classification bounds with items generated according to the 2PL IRT model with $J = 1,500$ items generated from a bank such that $c = 0$ and b normally distributed with a mean of 0 and a standard deviation of 0.707.

Fisher information at the ability estimate or the classification bound. Therefore, most of the algorithms would result in similar items selected regardless of the objective function. However, even though the algorithms result in more similar test lengths, the ELR still performs better than almost all of the other item selection algorithms across all classification bounds and both item banks. Only FI-middle ever results in shorter tests than the ELR and only for banks without lower asymptotes.

Finally, Figures 9 and 10 display the misclassification rates for those conditions represented by Figures 7 and 8. Unsurprisingly, the same pattern persists in Figures 9 and 10 as in Figures 5 and 6. That is to say, none of the misclassification rate graphs display

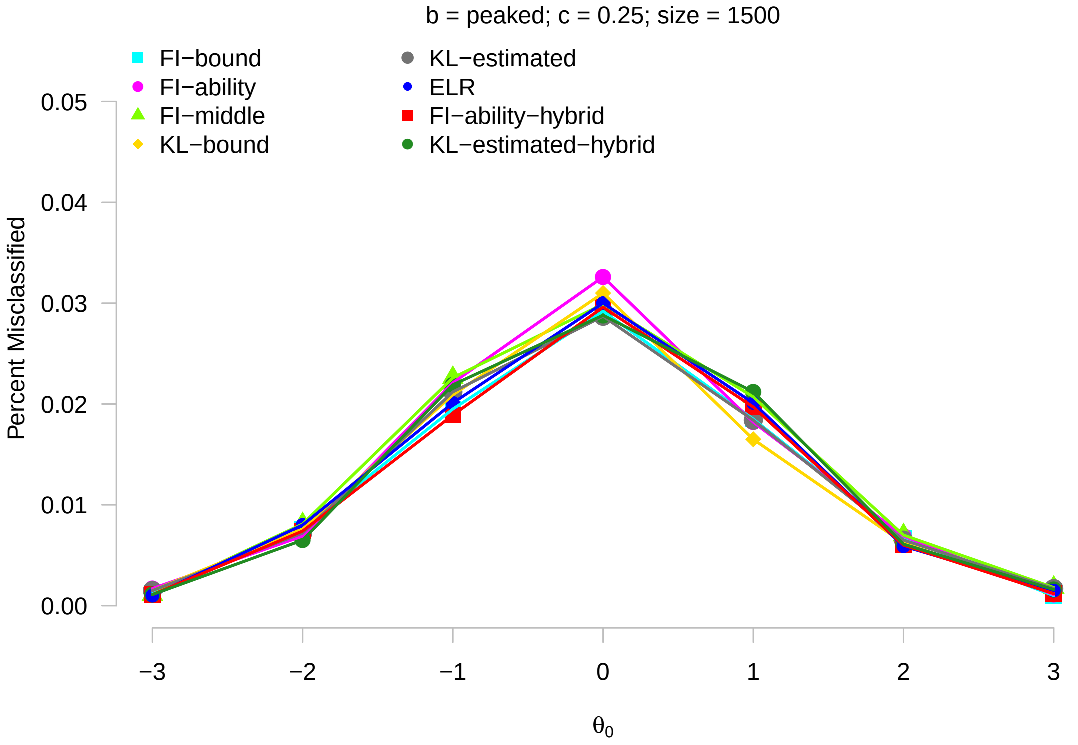

Figure 9. Percent misclassified using an SPRT stopping rule with item selection algorithms and classification bounds with items generated according to the 3PL IRT model with $J = 1,500$ items generated from a bank such that $c = .25$ and b normally distributed with a mean of 0 and a standard deviation of 0.707.

any noticeable pattern.

Of course, aggregate test lengths and misclassification rates only present part of the pictures. One typically cares less about whether tests are typically shorter for all candidates than whether tests are demonstrably shorter for any specific groups of candidates. In the next section, we examine the test lengths for all of the item selection bounds conditional on particular levels of θ. In this manner, we can determine whether the ELR item selection algorithm gains advantage with candidates close to the cut-point, candidates far from the cut-point, or across all ability levels.

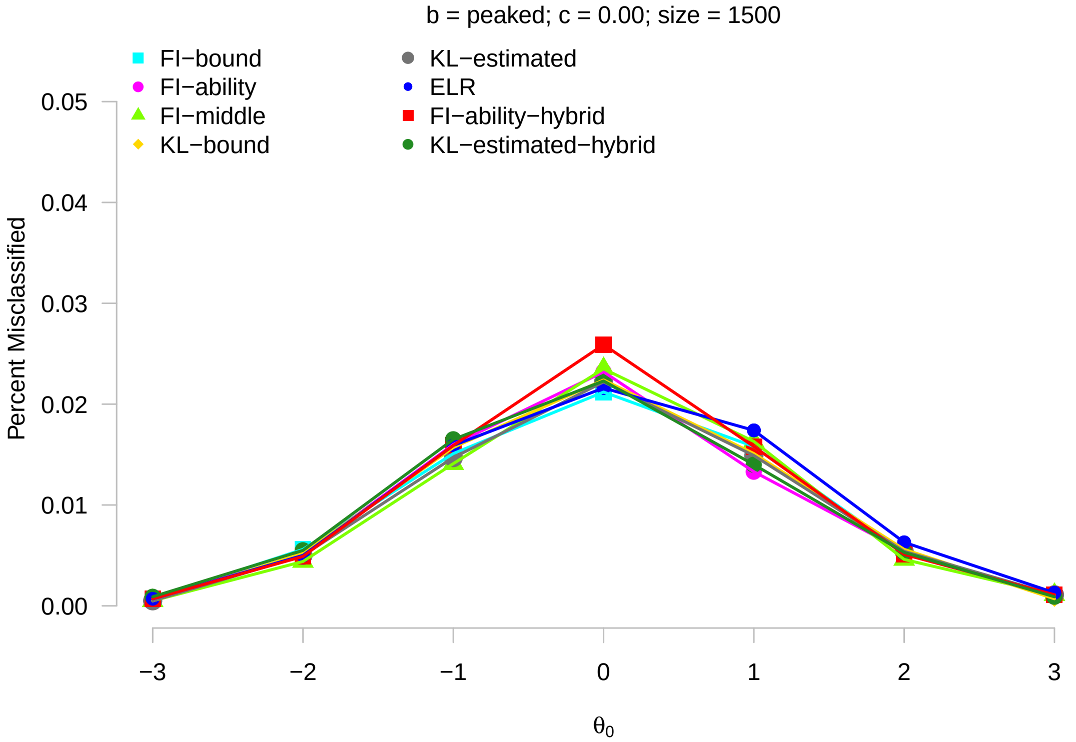

Figure 10. Percent misclassified using an SPRT stopping rule with item selection algorithms and classification bounds with items generated according to the 2PL IRT model with $J = 1,500$ items generated from a bank such that $c = 0$ and b normally distributed with a mean of 0 and a standard deviation of 0.707.

Conditional on Simulees with Similar θ

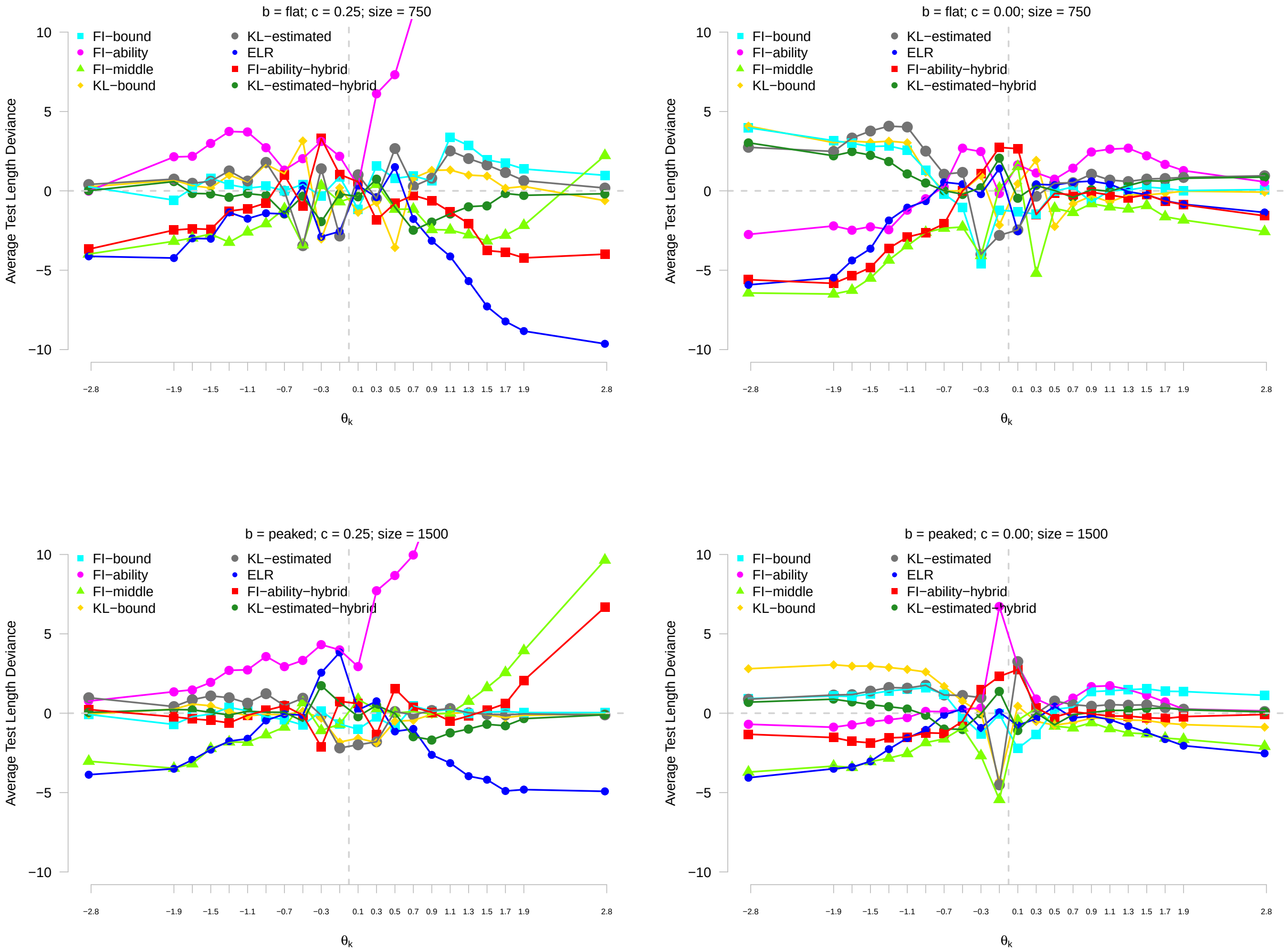

Figure 11. Average test length deviance using an SPRT stopping rule given various item selection algorithms with items generated from four different populations and a classification bound of $\theta_0 = 0$. In the upper quadrants, items were generated with $J = 750$ and b uniformly distributed between -4 and 4, and in the lower quadrants, items were generated with $J = 1,500$ and b normally distributed with a mean of 0 and a standard deviation of 0.707. In the left quadrants, items were generated such that $c = .25$, and in the right quadrants, items were generated such that $c = 0$. In all cases, test length was aggregated across simulees within an ability range of .2

Figure 11 presents the average test length deviance conditional on various item selection algorithm for the same four item banks described in the previous section but with a few caveats. First, in all cases, the classification bound represented by Figure 11 is $\theta_0 = 0$, which is displayed by a dashed grey vertical line in the center of each plot. Second, results are aggregated across all simulees with true ability within a limited range. The cut-points of that range are entries in the following set: $\{\min(\theta_i), -2, -1.8, -1.6, \ldots, -0.2, 0, 0.2, \ldots 1.6, 1.8, 2, \max(\theta_i)\}$. These cut-points were chosen to be large enough for averages to be calculated from a sufficient sample size (between $N = 125$ and $N = 800$) yet small enough for conclusions to be drawn about relatively narrow true ability regions.

All four quadrants of Figure 11 contain a fairly messy set of points. Lines interweave and criss cross frequently. However, some patterns are consistent across all four plots. First, FI-bound, which is often promoted as the optimal item selection algorithm for classification tests, rarely results in the shortest or near shortest tests. Only for true ability near the cut point does FI-bound outperform or almost outperform the other item selection algorithms in terms of test length. More often, FI-bound, as represented by the light blue squares, is at the median of item selection performance. Second, FI-ability performs poorly for most simulees given an item bank with $c = .25$. Yet, as predicted based on Nydick (2014), FI-ability results in much longer tests if $\theta_i > \theta_0$ than if $\theta_i < \theta_0$. Notice that ability need not even be far above the cut-point for FI-ability to result in much longer tests than the other item selection algorithms. Moreover, item selection algorithms based on choosing items to maximize Fisher information given an ability estimate (such as FI-ability-hybrid and FI-middle) also begin to result in longer tests if $\theta_i > \theta_0$ and $c = .25$. Both the light green triangle line and the red square line start to turn upward for high ability simulees in the left quadrants.

The ELR item selection algorithm does not always seem to perform better than all (or even most) item selection algorithms despite the overall conclusions drawn from the previous section. For low ability simulees if $c = .25$, and for low and high ability simulees if $c = 0$, FI-middle often results in shorter tests than the ELR, although the differences between those conditions are typically very small. Moreover, if $|\theta_i - \theta_0|$ is small, then the ELR often performs worse than standard item selection algorithms, such as FI-bound or KL-bound. One could argue that if $\theta_i = \theta_0 + 0.1$, then KL-bound would actually be the optimal item selection algorithm because then KL-bound would select items according to the ELR with the expectation taken with respect to θ_i rather than $\hat{\theta}_i$. The ELR should always perform worse than algorithms that inadvertently know true ability, as the ELR must always estimate θ_i. Therefore, one should not be shocked if KL-bound performs optimally or nearly optimally for $\theta_i \approx \theta_0 + \delta$. In nearly all of the quadrants, KL-bound, as depicted by the yellow diamond, results in the shortest test for θ_i slightly above $\theta_0 = 0$. Yet the ELR results in nearly the shortest tests for θ_i far above or far below the cut-point. In all four quadrants, the dark blue circle line trends downward for more extreme θ intervals. Presumably, for these simulees, the precision of the actual ability estimate is less important than ability being within a reasonable range of truth. This observation will return when proposing alternative item selection algorithms, based on the ELR, that take into consideration the uncertainty of the ability estimate more subtly than simply waiting until ability is sufficiently stable before using the ELR (such as FI-ability-hybrid).

Conclusion

Psychometricians continue to recommend selecting items for mastery tests at the cut-point separating categories (e.g., Huebner & Li, 2012). This recommendation has been made regardless of stopping rule used and without providing but a passing reference to other researchers (such as Spray & Reckase, 1994, as cited in each of Finkelman, 2008, or van Groen, Eggen, & Veldkamp, 2014, or Thompson, 2009). In Nydick (2014), we pointed out that even when using stopping rules that ignore $\hat{\theta}_i$ (such as the SPRT), "some relationship always exists between item responses, item parameters, and the log-likelihood ratio, so that estimates of ability ultimately anticipate future item responses and, thus, aid classification." (p. 223). The ELR item selection algorithm was designed to approximate how examinees might respond to the next item and to ensure that the item selected would optimize the information available in this expected response. And for most of the conditions described above, the ELR was either the most optimal or nearly the most optimal algorithm in terms of average test length. Only for simulees near the cut-score or if the item response model did not include a lower asymptote did other algorithms result in shorter tests than the ELR. And even in the latter case, if $c = 0$ for all items, the only algorithm that outperformed the ELR was an algorithm that was derived from the ELR for this special case.

Although we demonstrated that the ELR essentially reduces to FI-middle in the case of $c = 0$, one might still want to verify how items are chosen given different item selection algorithms. To this end, a small simulation was run to show the location of the item difficulty parameter (b) under four different item selection algorithms for models with $c = 0$ or $c = .25$ and for persons above and below the cut-point of $\theta_0 = 0$. Items were selected from a very large bank of $J = 10,000$ items, all with $a = 1$ and c dependent on the particular condition. The reason for setting $a = 1$ is because the item that maximizes Fisher information depends on both the location of the item (b) as well as the relationship between the item response and the underlying trait (a). By setting $a = 1$, we simply remove the nuisance variance of the latter and can better determine where items are being selected under different rules. The item bank was constructed to be very large so that a nearly optimal item could always be chosen regardless of rule or $\hat{\theta}$.

Figures 12 and 13 indicate the location of item selection under four of the item selection algorithms if $c = 0$ and $\theta_i = -2$ (Figure 12) or $\theta_i = 2$ (Figure 13). In all graphs, the cut-point is represented by a dashed grey line, and true ability is indicated by a dotted blue line. The person ability estimate after each item in the CAT, $\hat{\theta}_i$, is represented by a golden line, and the corresponding item difficulty (of the next item chosen), b, is depicted as a purple line. Note that Figures 12 and 13 depict items as one would expect. If using FI-bound, the location of b is always along the cut-point (the variability of b for larger j

only due to fewer available items close to that location). If using FI-ability, the location of b tracks the ability estimate. And if using FI-middle or the ELR, the location of b is essentially halfway between the classification bound and the ability estimate. The latter point, that the b-parameter that optimizes the ELR was close to halfway between θ_i and θ_0, was part of the justification for using FI-middle as an item selection algorithm in cases where $c = 0$.

Figures 14 and 15 indicate the location of item selection under the four algorithms described above if $c = .25$ for all items. For three of the four algorithms, the relative location of b does not depend on whether $\theta_i > 0$ or $\theta_i < 0$. FI-bound always selects items with b-parameter slightly below the classification bound, FI-ability always selects items with b-parameter slightly below the current ability estimate, and FI-middle always selects items slightly below halfway between $\hat{\theta}_i$ and θ_0. The reason that all of these algorithms select items slightly below their target θ is due to properties of the Fisher information function if $c > 0$ (e.g., Chang & Ying, 2009, Equation 4.2, p. 1478). In contrast to the standard algorithms, the ELR item selection location depends on whether $\theta_i > 0$ or $\theta_i < 0$. If $\theta_i < 0$, then the ELR behaves similarly to FI-middle, as shown in the bottom two quadrants of Figure 14 and as demonstrated based on the average test lengths depicted in the previous section. However, if $\theta_i > 0$, then the ELR adjusts the location of item selection to better account for how item difficulties affect the SPRT statistic. Without a lower asymptote or if $\theta_i < \theta_0$, the optimal item would be selected to best confirm our suspicions, that is, to be as difficult as possible while still having a high probability of being answered correctly (if $\hat{\theta}_i > \theta_0$) or as easy as possible while still having a high probability of being answered incorrectly (if $\hat{\theta}_i < \theta_0$). But if $c > 0$, then items selected with $b >> \theta_0$ will result in a likelihood ratio close to 1 regardless of item response. Therefore, the optimal $\hat{\theta}_i > \theta_0$ item for models with $c > 0$ would be difficult enough to challenge the current candidate while also being easy enough so that a candidate at $\theta_i = \theta_0 + \delta$ would have a non-negligibly larger than guessing change of answering the item correctly.

Based on the previous section, the ELR item selection algorithm did not result in the shortest test length in several scenarios, including if $\theta_0 = 0$. Note that if $\theta_0 = 0$, then the typical examinee would be close to the classification bound. In this case, algorithms that select items close to the classification bound without considering the current ability estimate (such as FI-bound or the KL-based algorithms) would be selecting items almost optimally for most of the simulees. The ELR, which requires an estimate of ability, could include an estimate of $\hat{\theta}_i$ far from θ_i early in a test, which could lead the ELR to be less than optimal. As was shown when presenting results conditional on a particular range of θ_i, classification-bound-based algorithms, such as FI-bound or KL-bound, only outperformed the ELR when $|\theta_i - \theta_0| \approx 0$. For these examinees, selecting items at the classification bound would be a better approximation to the ELR than the ELR itself. We tried to include algorithms that considered the uncertainty of the ability estimate early in a classification test, such as initially selecting items by maximizing Fisher information close to the ability estimate and then switching to the ELR once the ability estimate was reasonably precise. Unfortunately, this method is somewhat ad hoc, in that the SEM required before the algorithm would transition to the ELR was arbitrarily selected, and any benefits of the ELR would not commence until the transition. One could directly consider the uncertainty on $\hat{\theta}_i$ by maximizing the posterior ELR (e.g., Veerkamp & Berger, 1997, which they applied to Fisher information at $\hat{\theta}_i$ in the case of a precision CAT),

$$\mathrm{ELR}_j(\theta|\pi(\theta|\mathbf{y}_{i,j-1})) = \int_{\Theta} \pi(\theta|\mathbf{y}_{i,j-1})\mathrm{ELR}_j(\theta)d\theta, \tag{11}$$

where ELR_j is defined by Equation (7), $\pi(\theta|\mathbf{y}_{i,j-1})$ is the posterior distribution of θ given responses by person i to the first $j-1$ items on the CAT, and Θ is the region of integration. Maximizing the posterior ELR could reduce the inefficiency in the ELR for examinees with true ability close to the cut-score by limiting the effect of random noise in the ability estimate on the items selected early in a test.

The current research is limited by the narrow set of IRT models, decision rules, and number of cut-points. One could imagine that the optimality of the ELR (or even FI-middle) would generalize to unidimensional, polytomous IRT models, such as the graded response model (Samejima, 1996) or the generalized partial credit model (Muraki, 1996). However, only future research could demonstrate the generalizability of the derivations (from Nydick, 2014) or simulations (in this paper) to alternate IRT models. Moreover, many stopping rules exist for CCT, including the Generalized Likelihood Ratio (GLR; Thompson, 2009), the SPRT with Stochastic Curtailment (SCSPRT; Finkelman, 2008), the GLR with Stochastic Curtailment (SCGLR; Huebner & Fina, 2014), the Confidence Interval method (CI; Kingsbury & Weiss), and variations on Bayesian decision rules (e.g., Lewis & Shehan, 1990). Several of the algorithms (such as the GLR) are based on the likelihood ratio, so generalizing the ELR to these methods would be a straightforward extension of the original equations. However, other methods have different objective functions and would require new theory and alternate, optimal item selection algorithms. The point of this research is not to promote the ELR as the best item selection algorithm for variable length classification tests. The purpose of this paper is, instead, to argue that if maximal efficiency is paramount when designing an adaptive test, then an item selection algorithm should be chosen with the stopping rule in mind.

Many researchers still select items in classification testing by maximizing information at the classification bound (e.g., Smits & Finkelman, 2013; Haring, 2014, pp. 46, 84). Recently, Wyse and Babcock (2016) demonstrated that selecting items by maximizing information at the cut-point does not always maximize classification accuracy for fixed-length adaptive tests. In their study, they examined group level results across various item banks and examinee distributions. Only if the examinee ability distribution had a mean close to the cut-score or as the test length became large did the optimal item selection location become approximately at the classification bound. For most other cases, the optimal location to select items depended on the examinee distribution. If $\bar{\theta} > \theta_0$, items should typically be selected above the cut-score, and if $\bar{\theta} < \theta_0$, then items should typically be selected below the cut score. Although their study adopted a fixed length test rather than a variable length test using the SPRT, varied the distribution of simulees rather than the classification bound, and examined the classification accuracy/consistency rather than the decision efficiency, they still arrived at a similar conclusion to the one in this paper. That is, information about a simulee's location relative to a classification bound depends on both the simulee as well as the bound. Ignoring the simulee or the classification bound when choosing an item limits being able to quantify the knowledge one might obtain from the next item, and, thus, reduces the potential efficiency of the test.

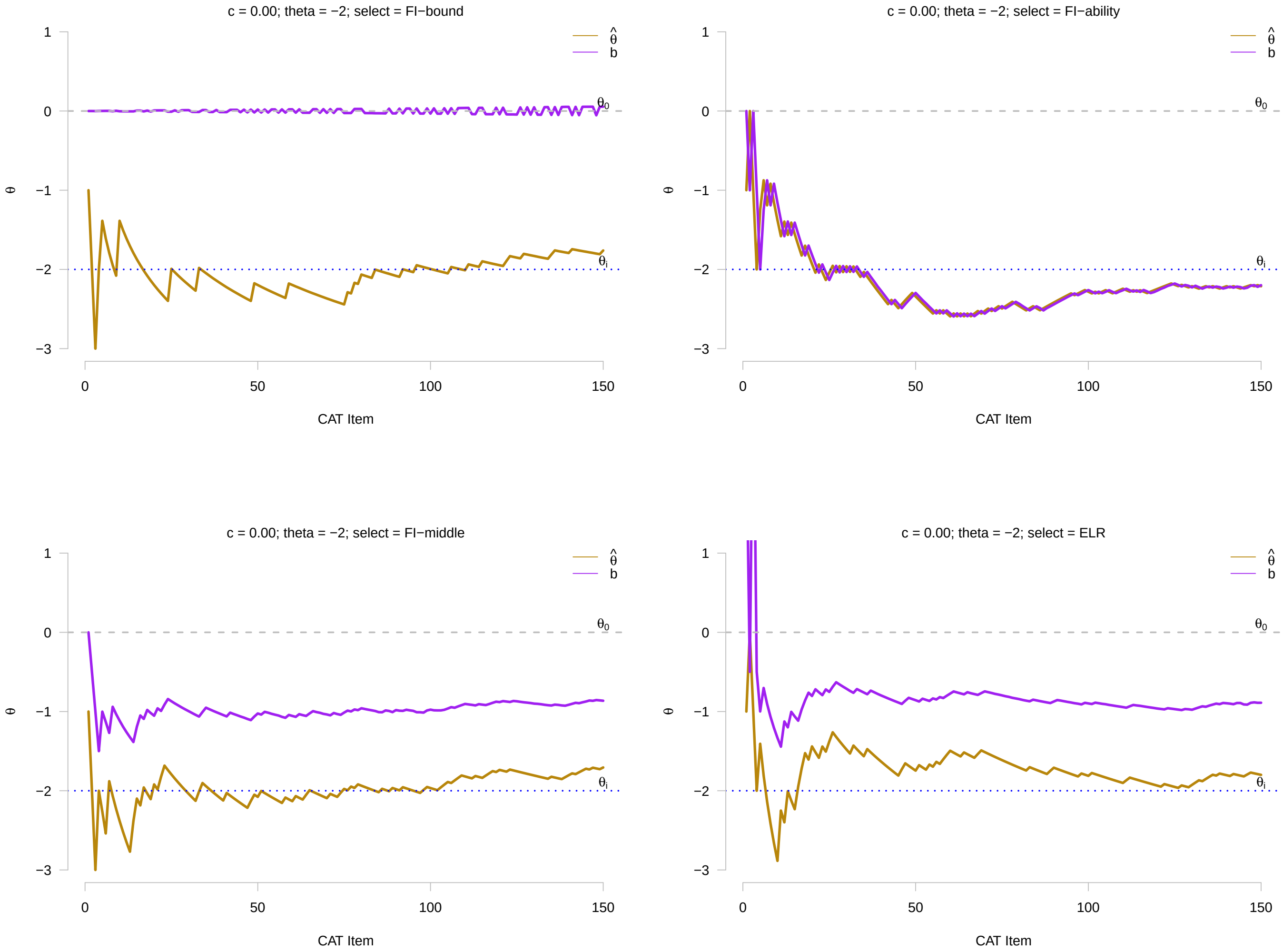

Figure 12. Item difficulty (b-parameter) selected after each item given a fixed length test with $j_{\max} = 150$ and an item bank with $a = 1$ and $c = 0$ for all items. The true person parameter, θ_i, was set to -2, and the classification bound, θ_0, was set to 0. The upper left and right quadrants are FI-bound and FI-middle, and the lower left and right quadrants are FI-middle and the ELR algorithm.

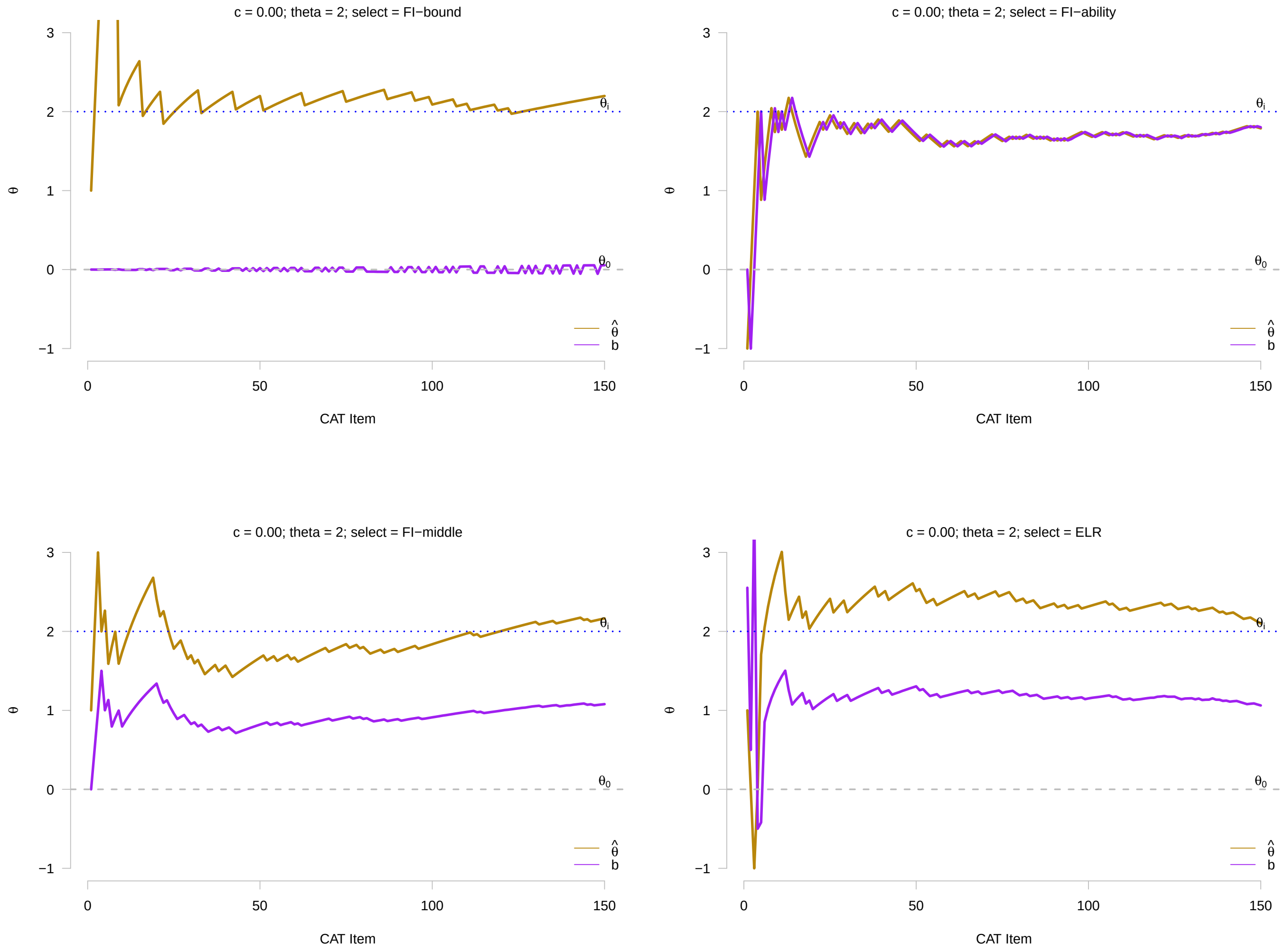

Figure 13. Item difficulty (b-parameter) selected after each item given a fixed length test with $j_{\max} = 150$ and an item bank with $a = 1$ and $c = 0$ for all items. The true person parameter, θ_i, was set to 2, and the classification bound, θ_0, was set to 0. The upper left and right quadrants are FI-bound and FI-middle, and the lower left and right quadrants are FI-middle and the ELR algorithm.

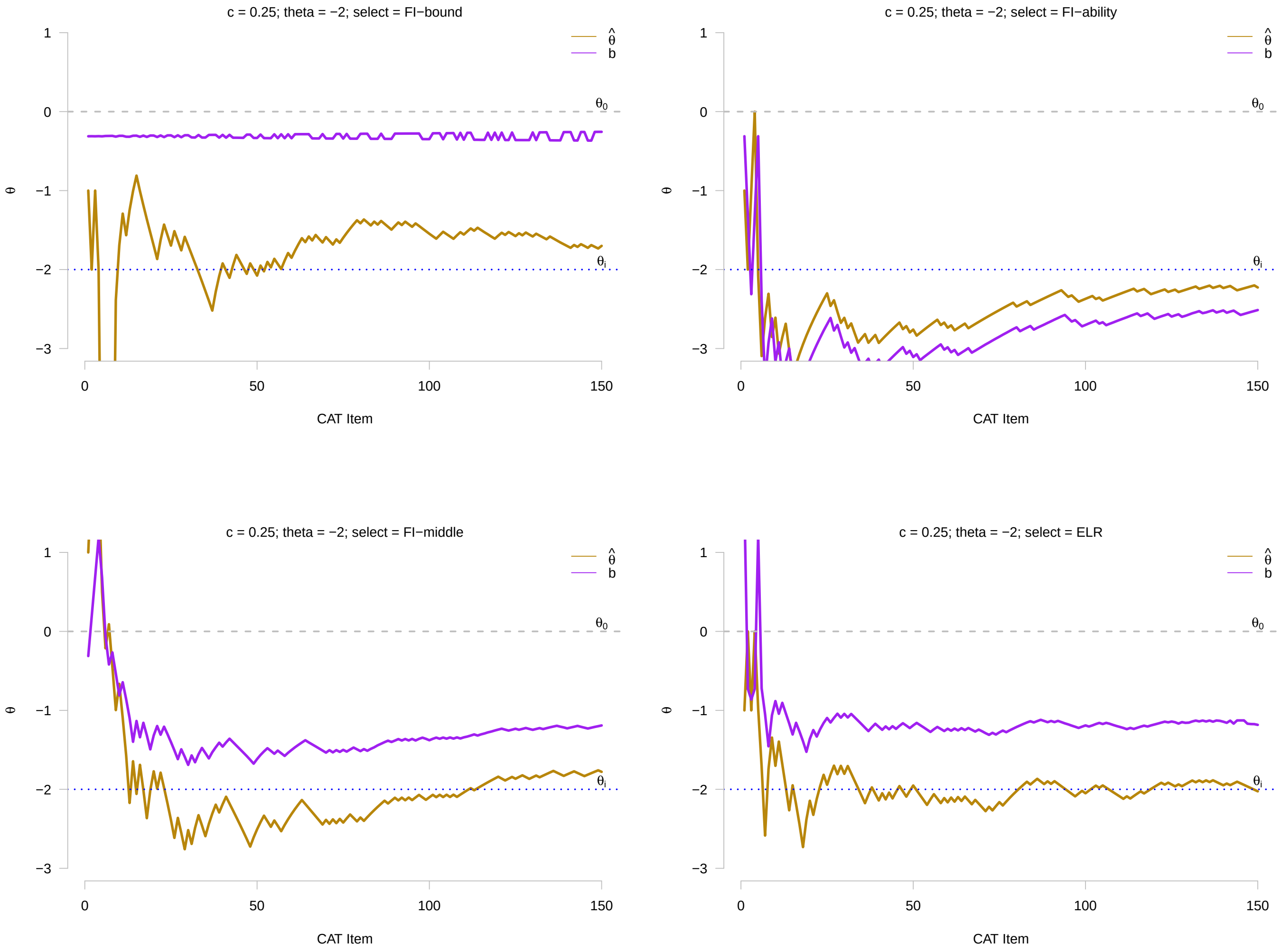

Figure 14. Item difficulty (b-parameter) selected after each item given a fixed length test with $j_{\max} = 150$ and an item bank with $a = 1$ and $c = .25$ for all items. The true person parameter, θ_i, was set to -2, and the classification bound, θ_0, was set to 0. The upper left and right quadrants are FI-bound and FI-middle, and the lower left and right quadrants are FI-middle and the ELR algorithm.

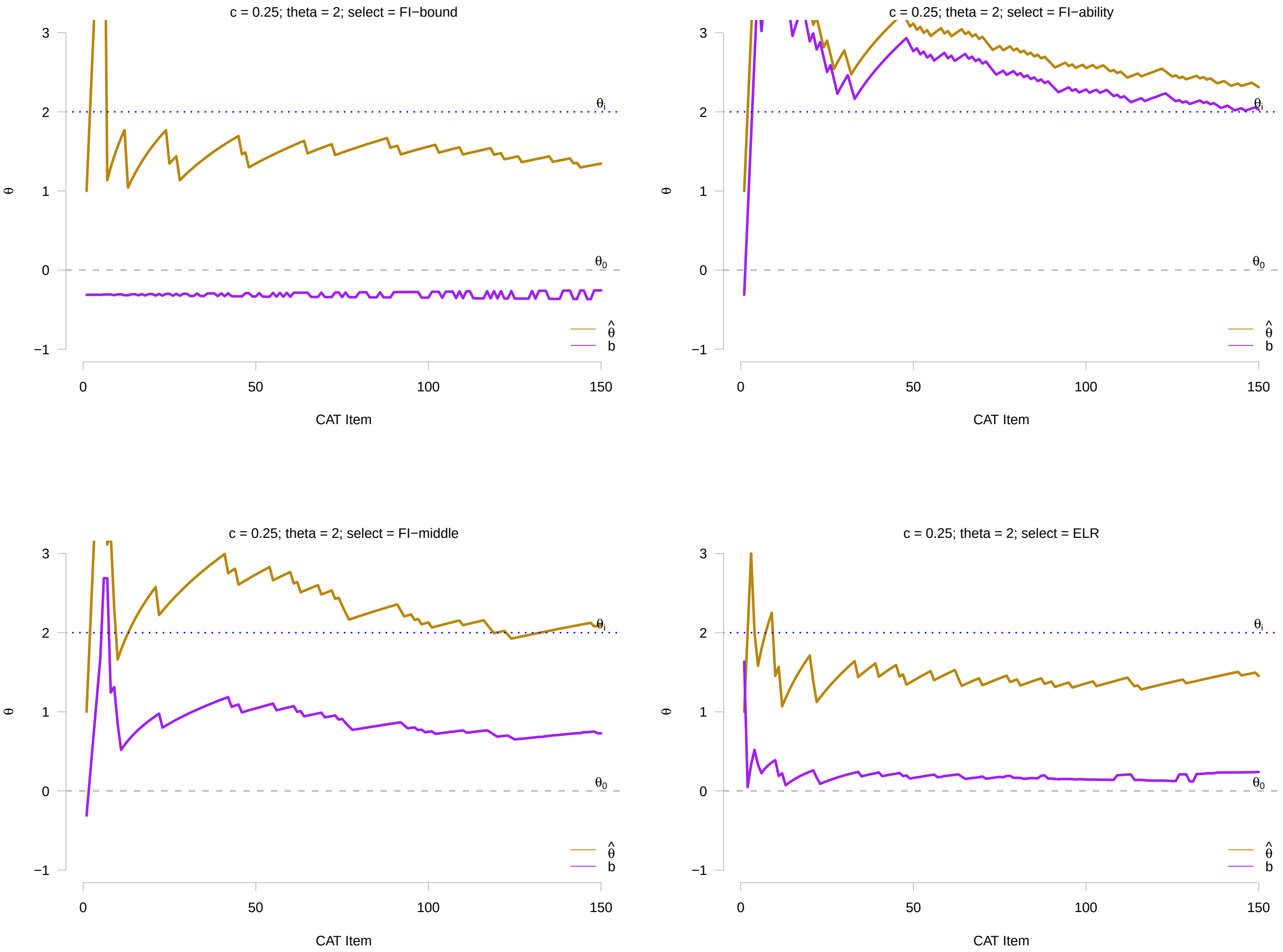

Figure 15. Item difficulty (b-parameter) selected after each item given a fixed length test with $j_{\max} = 150$ and an item bank with $a = 1$ and $c = .25$ for all items. The true person parameter, θ_i, was set to 2, and the classification bound, θ_0, was set to 0. The upper left and right quadrants are FI-bound and FI-middle, and the lower left and right quadrants are FI-middle and the ELR algorithm.

References

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

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

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, *65*, 457–482.

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

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

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

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

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

Haring, S. H. (2014). A comparison of three statistical testing procedures for computerized classification testing with multiple cutscores and item selection methods (Doctoral dissertation). Retrieved from https://repositories.lib.utexas.edu/bitstream/handle/2152/24838/HARING-DISSERTATION-2014.pdf

Huebner, A. (2012). Item overexposure in computerized classification tests using sequential item selection. *Practical Assessment, Research & Evaluation*, *17*, 1–9.

Huebner, A., & Li, Z. (2012). A stochastic method for balancing item exposure rates in computerized classification tests. *Applied Psychological Measurement*, *36*, 181–188.

Huebner, A. R., & Fina, A. D. (2014). The stochastically curtailed generalized likelihood ratio: A new termination criterion for variable-length computerized classification tests. *Behavior Research Methods*, *47*, 549–561.

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.* New York: Academic Press

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

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.

Muraki, E. (1996). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-164). New York: Springer.

Nydick, S. W. (2014). The sequential probability ratio test and binary item response models. *Journal of Educational and Behavioral Statistics*, *39*, 203–230.

Samejima, F. (1996). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

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

Smits, N., & Finkelman, M. D. (2013). A comparison of computerized classification testing and computerized adaptive testing in clinical psychology. *Journal of Computerized Adaptive Testing*, *1*, 19–37.

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.

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

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.

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

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

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

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.

Wyse, A. E., & Babcock, B. (2016). Does maximizing information at the cut score always maximize classification accuracy and consistency? *Journal of Educational Measurement*, *53*, 23–44.