Psychological Test and Assessment Modeling, Volume 54, 2012 (4), 432-449

Capitalization on chance in variable-length classification tests employing the Sequential Probability Ratio Test

Jeffrey M. Patton[1], Ying Cheng[2], Ke-Hai Yuan[2] & Qi Diao[3]

Abstract

The sequential probability ratio test (SPRT) is a popular termination criterion for variable-length classification tests. The SPRT is often paired with cut-based item selection in which item information is maximized at the cut point. However, items are chosen on the basis of their parameter estimates, and capitalization on chance may occur. We investigated the effects of capitalization on chance on test length and classification accuracy in several variable-length test simulations. In addition to capitalizing on large discrimination estimates, the item selection criterion chose items with difficulty estimates systematically higher or lower than their true difficulty values. This capitalization on chance had non-negligible effects on both test length and classification accuracy and induced an inverse relationship between them, though the particular effects were highly sensitive to the cut location. The results also indicate that implementing item exposure control effectively reduced the effects of capitalization on chance on testing outcomes.

Key words: sequential probability ratio test, classification testing, variable-length testing, capitalization on chance, item calibration error

[1] *Correspondence concerning this article should be addressed to:* Jeffrey M. Patton, M.Ed., University of Notre Dame, 209 Haggar Hall, Notre Dame, IN 46556, USA; email: jpatton1@nd.edu

[2] University of Notre Dame

[3] CTB/McGraw-Hill

Introduction

Variable-length computerized classification tests (VL-CCTs) are designed to accurately classify examinees using as few items as possible (Thompson, 2007). As the name implies, the length of the test is allowed to vary across examinees, and a rule is needed to determine when to stop the test. A popular termination criterion in the item response theory (IRT) framework is the sequential probability ratio test (SPRT; Reckase, 1983; Wald, 1947). The SPRT formulates the classification decision as a hypothesis-testing problem using a likelihood ratio test, and a major advantage is that estimation of the latent trait is not required. To make classifications efficiently, an appropriate item selection criterion is also needed. In computerized adaptive testing (CAT), the goal of item selection is to optimize the match between items and the current estimate of the latent trait. In classification testing, however, the goal is to decide whether an examinee's value of the latent trait is above or below a predetermined cut point. Thus it has been argued that when paired with the SPRT, item selection should seek to optimize the match between items and the cut point (Thompson, 2009).

In practice, only estimates of the item parameters are available, and because item selection involves optimization, capitalization on chance may occur. This phenomenon was demonstrated by van der Linden and Glas (2000) in the context of fixed-length CAT: several popular item selection criteria tended to choose items with spuriously high discrimination parameter estimates, which had detrimental effects on item exposure rates and latent trait estimates. Concerning VL-CCTs employing the SPRT and cut-based item selection, Spray and Reckase (1987) and Kalohn and Spray (1999) investigated the effects of item calibration error on test length and classification accuracy. Neither of these studies found adverse effects, but for reasons to be elaborated below, we believe that the relationship between calibration error and the SPRT deserves further investigation. Our goal was to conduct a more comprehensive simulation study to examine the effects of various magnitudes of item calibration error on practical outcomes of VL-CCTs employing the SPRT.

The rest of the paper is organized as follows. First we review the SPRT procedure as implemented in VL-CCTs and research concerning item selection and capitalization on calibration error. Next we hypothesize the effects of calibration error on test length and classification accuracy. Lastly, we present the results of a simulation to demonstrate the effects of calibration error on VL-CCTs, followed by a discussion of practical implications.

The Sequential Probability Ratio Test

As stated above, the SPRT formulates the classification decision as a hypothesis-testing problem (Spray & Reckase, 1996). First, in a two-category situation, a cut point θ_0 is chosen which divides the θ scale into ordered categories (e.g., pass and fail). For a given examinee, the goal is to decide between the composite hypotheses: $H_0 : \theta < \theta_0$ versus $H_1 : \theta \geq \theta_0$. The SPRT is implemented by instead considering two simple hypoth-

eses: $H_0 : \theta = \theta_1$ versus $H_1 : \theta = \theta_2$, where $\theta_1 < \theta_0 < \theta_2$. The interval (θ_1, θ_2) is called the indifference region, and it defines a range of ability values for which we accept misclassifications due to measurement error. Its endpoints are often chosen such that $\theta_1 = \theta_0 - \delta$ and $\theta_2 = \theta_0 + \delta$, where δ is a positive constant.

The hypothesis test is conducted using a likelihood ratio. Given an examinee's vector of responses ($\boldsymbol{u}$) to *m* binary items, the log-likelihood ratio statistic is computed:

$$\log(\lambda) = \log\left[\frac{L(\theta_2 \mid \boldsymbol{u})}{L(\theta_1 \mid \boldsymbol{u})}\right]. \tag{1}$$

This statistic is compared to upper and lower decision points (A and B, respectively), which are functions of the nominal Type I and Type II error rates (α and β, respectively), chosen prior to testing. The test ends when $\log(\lambda) > A = \log[(1-\beta)/\alpha]$ or $\log(\lambda) < B = \log[\beta/(1-\alpha)]$, in which case the examinee is classified as passing or failing, respectively. In practice, a maximum test length is specified. If this maximum is reached before $\log(\lambda)$ reaches either of the decision points, it may be compared to $C = (A+B)/2$ to make the decision (Spray & Reckase, 1996). In this way, the SPRT serves as both termination criterion and decision rule, and estimation of the latent trait is not required.[4]

A number of factors are known to influence test length under the SPRT. This is easy to show under conditions of local independence, in which case Equation 1 can be expressed as a sum of log-likelihood ratios associated with individual item responses:

$$\log(\lambda) = \sum_{j=1}^{m} \log(\lambda_j) = \sum_{j=1}^{m} \left[\log \frac{P_j(\theta_2)^{u_j} Q_j(\theta_2)^{1-u_j}}{P_j(\theta_1)^{u_j} Q_j(\theta_1)^{1-u_j}}\right], \tag{2}$$

where u_j is the j^{th} item response, $P_j(\theta)$ is the probability of a correct response, and $Q_j(\theta) = 1 - P_j(\theta)$. At the beginning of the test, $\log(\lambda)$ is equal to zero. As the test proceeds, correct responses "push" $\log(\lambda)$ in the positive direction and incorrect responses push it in the negative direction. One important influence on test length is the width of the indifference region. A narrow region (i.e., a small value for δ) implies that θ_1 and θ_2 are close together. Thus the value of $|\log(\lambda_j)|$ will be small for each item, and a relatively large number of items will be required to make a decision. Another important factor is the true value of the latent trait. When items are chosen to match the cut point, tests tend to be short for examinees with θ values far from the cut, whereas examinees near the cut (e.g., in the indifference region) have much longer tests (Spray & Reckase, 1996). The reason is that examinees near the cut tend to have a roughly even mix of correct and incorrect responses. As a result, $\log(\lambda)$ will alternate between positive and negative values, rather than move quickly toward either of the decision points. Thus more observations are required before a classification decision can be confidently made.

[4] Because a maximum test length is imposed, this procedure is technically referred to as the truncated SPRT (Finkelman, 2008).

In contrast, examinees far from the cut tend to have much more homogeneous response patterns, resulting in much shorter tests. Lastly, the properties of the items selected for administration have important implications for test length. We discuss these properties in the next section.

Item selection and capitalization on chance

As stated above, $\log(\lambda)$ is equal to zero at the beginning of the test, so the goal of item selection is to choose items such that $\log(\lambda)$ moves away from zero toward either of the decision points as quickly as possible (Lin & Spray, 2000). A popular method to achieve this goal is to maximize Kullback-Leibler (KL) information at the bounds of the indifference region (Eggen, 1999; Lin & Spray, 2000; Thompson, 2009). When paired with the SPRT, KL information for item *j* is usually computed as follows:

$$K_j\left(\theta_2 \,||\, \theta_1\right) = E_{\theta_2} \log\left[\frac{L(\theta_2 \,|\, u_j)}{L(\theta_1 \,|\, u_j)}\right], \tag{3}$$

where the expectation is taken with respect to θ_2 (Eggen, 1999).[5] This formula shows that KL information is a measure of the distance between two distributions: the distribution of an item response given θ_2 versus that given θ_1. In other words, it expresses the expected information in the observed data for distinguishing between the two associated hypotheses (i.e., $\theta = \theta_2$ versus $\theta = \theta_1$, respectively). By sequentially choosing items to maximize KL information, the values of $\left|\log(\lambda_j)\right|$ are maximized, thus minimizing the number of items needed to reach a decision.

Of course, KL information depends on the values of the item parameters. If the probability of a correct response is modeled by the two-parameter logistic model (2PLM):

$$P_j(\theta) = \frac{\exp\left[a_j\left(\theta - b_j\right)\right]}{1 + \exp\left[a_j\left(\theta - b_j\right)\right]}, \tag{4}$$

where a_j and b_j are item discrimination and difficulty, respectively, Equation 3 can be computed by

$$K_j\left(\theta_2 \,||\, \theta_1\right) = a_j\left(\theta_2 - \theta_1\right) P_j\left(\theta_2\right) + \log\left[\frac{Q_j(\theta_2)}{Q_j(\theta_1)}\right] \tag{5}$$

(Eggen, 1999). Using a cut of $\theta_0 = 0$ with $\delta = 0.25$ (so that $\theta_1 = -0.25$ and $\theta_2 = 0.25$), Figure 1 displays the value of KL information attained for different combinations of item

[5] The expectation can instead be taken with respect to θ_1, but note that $K_j\left(\theta_2 \,||\, \theta_1\right) \neq K_j\left(\theta_1 \,||\, \theta_2\right)$. In practice, the expectation is usually taken with respect to θ_2.

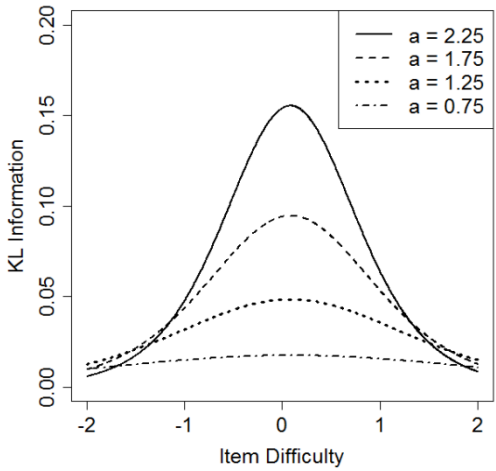

Figure 1:
Value of Kullback-Leibler Information for Different Combinations of Item Difficulty and Discrimination

difficulty and discrimination. For a fixed value of *b*, KL information is a unimodal function of *a*. As shown in Figure 1, however, when *b* is close to the cut, information is increasing in *a* for the range of *a* values typically encountered in practice. For a given value of *a*, the highest value of KL information is attained when *b* is located slightly to the right of the cut. For these items, the optimal *b* value is close to 0.085 (it is not exactly at zero because the expectation in Equation 3 is taken with respect to θ_2).

However, the true item parameter values are never known in practice, so items are chosen on the basis of their parameter estimates. Because optimal values of KL information depend on optimal values of the item parameter estimates, capitalization on chance may occur. This phenomenon has been demonstrated with the maximum Fisher information criterion under adaptive item selection. Under the 2PLM, Fisher information for item *j* is computed by

$$I_j = a_j^2 P_j(\theta) Q_j(\theta) . \tag{6}$$

Similar to KL information, this criterion prefers items with large *a* values among those with *b* values close to θ. However, the largest *a* estimates in an item pool tend to be spuriously large: the sum of large true *a* values and large, positive calibration errors (van der Linden & Glas, 2000). Thus Fisher information evaluated with respect to item parameter estimates tends to be larger than that evaluated with respect to the true parameter values (Hambleton & Jones, 1994; Hambleton, Jones, & Rogers, 1993). In addition to its effect on Fisher information, capitalization on chance may have adverse effects on important testing outcomes: it may negatively affect θ recovery (van der Linden & Glas, 2000), yield spuriously low standard errors for maximum likelihood ability estimates

(Hambleton, Jones, & Rogers, 1993), and lead to highly skewed distributions of item exposure rates (van der Linden & Glas, 2000). This problem of capitalization on chance is exacerbated when item calibration errors tend to be large, for example, when the ratio of pretest sample size to the number of model parameters is small (Hambleton & Jones, 1994). The problem also depends on the ratio of test length to the conditional size of the item pool; when there are fewer items to choose from, it is less likely to systematically choose those items with spuriously large discrimination estimates (van der Linden & Glas, 2000).

When maximizing KL information about the cut point, will capitalization on chance occur? KL and Fisher information prefer similar types of items, so it is reasonable to expect KL information to also capitalize on spuriously large discrimination estimates. Under cut-based item selection, however, information is evaluated over a small region of ability. Thus the effective size of the pool is much smaller than that under adaptive item selection, and capitalization on chance may be reduced. The problem may be reduced further when exposure control is implemented. Under cut-based selection, a common method of exposure control is to introduce a random component (Lin & Spray, 2000). First, specify a bin depth d. The first d most informative items comprise the first bin, the next d most informative items comprise the second bin, and so on. As the test proceeds, an item is randomly selected from each bin, beginning with the most informative. As the test proceeds, items will eventually be chosen from among relatively uninformative bins (this is particularly true when the bin depth is large), and it is only in the most informative bins that we expect to find spuriously high a estimates.

What might be the practical effects of capitalization on chance when the SPRT is the termination criterion? If the chosen items have spuriously large a estimates, the value of KL information will be spuriously large. As stated above, large values of KL information produce large changes in the log-likelihood ratio statistic (Equation 2), so this should yield spuriously short tests. If a test is spuriously short, this implies that the classification decision was made prematurely, and so short tests may yield worse classifications.

Purpose of the study

To our knowledge, two studies have investigated the effects of item calibration error on the efficiency and accuracy of VL-CCTs employing the SPRT (viz., Kalohn & Spray, 1999; Spray & Reckase, 1987). Though the investigators concluded that the effects were inconsequential, these studies can be extended in at least two ways. First, each study employed item parameter estimates based on a single, large pretest sample. The size of calibration errors is inversely related to the sample size (on average); so if the pretest sample is sufficiently large, the effects of calibration error are probably negligible. However, applications of IRT models are increasingly common in areas such as psychological and health research which often lack the luxury of large samples. Thus we propose to use a range of pretest sample sizes. Second, the previous studies averaged results over an entire sample of examinees. As demonstrated by van der Linden and Glas (2000), the effects of capitalization on chance may strongly depend on the true θ value. Thus our

goal is to conduct simulations, similar in design to those of van der Linden and Glas (2000), to study the effects of calibration error on VL-CCTs employing the SPRT and cut-based item selection. In particular, we wish to manipulate the magnitude of calibration error via the pretest sample size, and examine the effects on average test length and classification accuracy conditional on several true values of θ.

Method

A pool of "true" 2PLM parameters for 540 items was obtained from a retired item pool of a large-scale achievement test. The pool was calibrated using pretest sample sizes of 2500, 1000, and 500, yielding three sets of item parameter estimates which differed with respect to the average magnitude of calibration errors. Each calibrated pool was then "administered" to a future scoring sample. To mimic reality, item selection and evaluation of the log-likelihood ratio were conducted with respect to item parameter estimates, whereas responses were generated using the true parameters of the corresponding items. We also included a baseline condition in which the true item parameters were used for item selection, response generation, and evaluation of the log-likelihood ratio. This corresponds to the hypothetical situation in which the calibration sample size is infinite (i.e., $N = \infty$) and the item parameter values are known exactly.

Construction of item pools

Descriptive statistics of the parameter values are shown in Table 1. The distribution of item difficulty is roughly symmetric with a mean of 0.15, and the distribution of discrimination has a mean of 1.03 and a slight positive skew. The scatterplot of discrimination versus difficulty values (Figure 2) show a medium positive correlation ($r = .48$).

Table 1:
Descriptive Statistics of True Item Parameter Values

parameter	mean	std. dev.	skewness
a	1.03	0.33	0.56
b	0.15	1.02	−0.21

Next, item parameter estimates were obtained for pretest sample sizes of $N = 2500$, 1000, and 500. Rather than generate response data and estimate the item parameters, parameter estimates for a given item were drawn from their asymptotic bivariate normal sampling distribution, a method that has been employed in similar investigations (Spray & Reckase, 1987; van der Linden & Glas, 2000). This was done to avoid convergence problems during item calibration. Ability in the pretest samples was assumed to follow a standard normal distribution, and difficulty in the item pool ranged from −3.11 to 3.28. Even with pretest samples as large as 500, 2PLM calibration with marginal maximum likelihood would be likely to encounter convergence problems under these conditions.

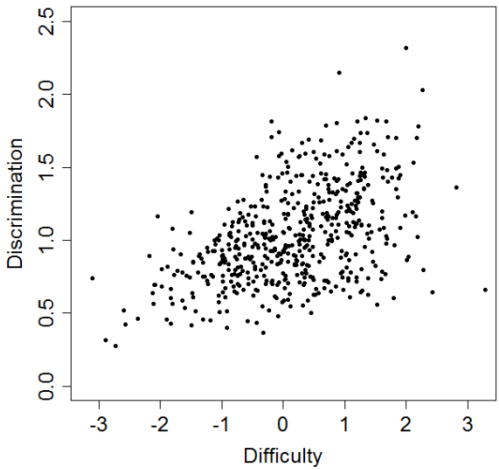

Figure 2:
Scatterplot of True Item Parameter Values (Discrimination versus Difficulty)

Item parameter estimates were generated as follows. First we computed the asymptotic covariance matrix $\boldsymbol{\Sigma}$ of maximum likelihood item parameter estimates given a (known) standard normal distribution of ability (see Thissen & Wainer, 1982). This method does not require observed data and uses Fisher information so that $\boldsymbol{\Sigma}$ is block-diagonal with all inter-item covariances equal to zero. Thus to simulate the calibration of a given item, a random sample was drawn from a bivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix from the appropriate 2×2 block in $\boldsymbol{\Sigma}$. These "errors" were then added to the true parameter values. To avoid dependence of the simulation results on a particular set of parameter estimates, 75 sets of estimates were generated for each value of *N*. Table 2 displays the mean squared error (MSE) of item parameter estimates for each combination of parameters and sample size. As expected, smaller values of *N* yielded larger deviations of estimates from their true values (on average).

Table 2:
Recovery of Item Parameters (Mean Squared Error) Averaged Across 75 Replications

N	a	b
2500	.004	.006
1000	.009	.015
500	.018	.031

Study 1

The simulated tests were designed to make a single pass/fail decision using the SPRT. A cut point of $\theta_0 = 1$ was chosen for all simulations because it is near the peak of pool information evaluated with respect to the true item parameters, and it corresponds to an expected proportion-correct score of .65. We also fixed the width of the indifference region ($\delta = 0.25$) and placed an equal weight on each type of error so that $\alpha = \beta = .05$. Once a minimum of five items was administered, testing continued until the log-likelihood ratio statistic (Equation 2) was beyond the upper or lower decision point ($A = 2.94$ or $B = -2.94$, respectively). If a maximum test length of 40 or 80 items was reached, the statistic was compared to zero (halfway between the upper and lower decision points), and a forced decision was made.

In all conditions, items were selected to maximize Kullback-Leibler information (Equation 5) at the cut point. Additionally, simulations were conducted with two bin depths ($d = 1$ or 5). Note that a bin depth of one is equivalent to having no exposure control. With a bin depth of five, an item is randomly selected from each bin of five items, and thus constitutes a condition with exposure control. Lastly, item pools associated with each pretest sample size were utilized within each test scenario. This resulted in three fully crossed factors: two maximum test lengths (40 or 80 items), two bin depths ($d = 1$ or 5), and four pretest sample sizes ($N = \infty$, 2500, 1000, or 500), yielding 16 conditions in all.

Study 2

We conducted additional simulations to examine the effects of cut location on test outcomes. Alternative cut locations of $\theta_0 = -1$ and $\theta_0 = 0.5$, corresponding to expected proportion-correct scores of .27 and .55, respectively, were arbitrarily selected. In these simulations, the maximum test length was fixed at 40 and no exposure control was used (i.e., $d = 1$). Thus only two factors were crossed: cut location ($\theta_0 = -1$ or 0.5) and pretest sample size ($N = \infty$, 2500, 1000, or 500). All other test features were identical to those in Study 1.

Dependent measures

For both Study 1 and Study 2, the scoring sample consisted of 300 examinees at each of 17 equally-spaced θ values between −2 and 2 in increments of 0.25. Each of the 75 calibrated item pools was administered to a separate scoring sample, and the dependent measures were averaged across the 75 replications. In this way, the outcomes at each θ value reflect two sources of error: random responses to items (i.e., measurement error) and capitalization on item calibration error via the item selection method. At each true θ value, we computed the average test length and classification accuracy, i.e., the percentage of accurately classified examinees.

Finally, all data generation, simulations, and analyses were performed in R (R Development Core Team, 2011) with codes written by the first author.

Results: Study 1

Effects of pretest sample size

We hypothesized that, due to capitalization on chance, tests would be spuriously short for small N at a fixed value of θ. For the condition with a maximum of 40 items and no exposure control, the top-left portion of Table 3 displays the average test length (ATL) for each value of N, as well as the difference between the $N = 500$ results and the baseline results (i.e., $N = 500$ results minus $N = \infty$ results). The effect of N was negligible at θ values far from the cut; this was because very few items were required to make a decision for these examinees, regardless of N. Thus results for these examinees are not shown. For those examinees near the cut, however, tests were spuriously short for small N below the cut, and tests were spuriously long for small N above the cut. However, the effect of N on test length is not large in an absolute sense. At $\theta = 1$, for example, tests under $N = 500$ were only 1.2 items shorter (on average) than tests under $N = \infty$. To confirm that these differences were not simply due to sampling error, we computed the standardized difference in ATL for each value of N relative to the baseline condition. At each θ value, mean differences were divided by the standard error of the baseline ATL (standard error = standard deviation of 75 ATLs / $\sqrt{75}$, where 75 is the number of replications). These results are shown in Figure 3a. Horizontal bars denote standardized differences of −2 and 2, and it is clear that for each combination of N and θ, differences are often much larger than that. For example, the absolute standardized differences under $N = 500$ range from 12 to 31 units.

Table 3:
Conditional Average Test Length ($\theta_0 = 1$)

		No Exposure Control ($d = 1$)							**Exposure Control ($d = 5$)**						
		θ Level							**θ Level**						
Max Length	N	0.25	0.50	0.75	1.00	1.25	1.50	1.75	0.25	0.50	0.75	1.00	1.25	1.50	1.75
40 Items	∞	13.6	19.2	27.8	32.9	27.3	17.9	12.2	18.3	25.1	32.7	36.0	32.6	24.5	17.3
	2500	13.3	18.6	27.2	32.6	27.6	18.2	12.3	18.1	24.8	32.4	35.9	32.6	25.0	17.6
	1000	13.3	18.4	26.7	32.2	27.7	18.7	12.5	17.7	24.3	32.1	35.6	32.9	25.2	17.8
	500	12.7	17.3	25.4	31.7	28.4	19.4	13.0	17.3	23.7	31.3	35.6	32.9	25.7	18.3
	Difference	**−0.9**	**−1.9**	**−2.4**	**−1.2**	**1.0**	**1.5**	**0.9**	**−1.0**	**−1.4**	**−1.4**	**−0.4**	**0.3**	**1.2**	**1.0**
80 Items	∞	13.8	20.1	36.2	50.2	35.4	18.8	12.1	18.9	30.7	50.6	62.9	51.1	30.7	18.3
	2500	13.4	19.6	34.8	50.3	36.0	19.2	12.3	18.7	29.9	50.1	63.0	51.5	31.2	18.6
	1000	13.3	19.0	33.6	49.1	36.9	19.8	12.5	18.2	29.4	49.5	62.7	52.1	31.3	18.7
	500	12.7	18.0	31.8	47.9	38.2	21.0	13.1	17.9	28.5	48.0	61.9	53.1	32.8	19.5
	Difference	**−1.2**	**−2.1**	**−4.4**	**−2.3**	**2.8**	**2.2**	**1.0**	**−1.0**	**−2.2**	**−2.6**	**−1.1**	**2.0**	**2.1**	**1.2**

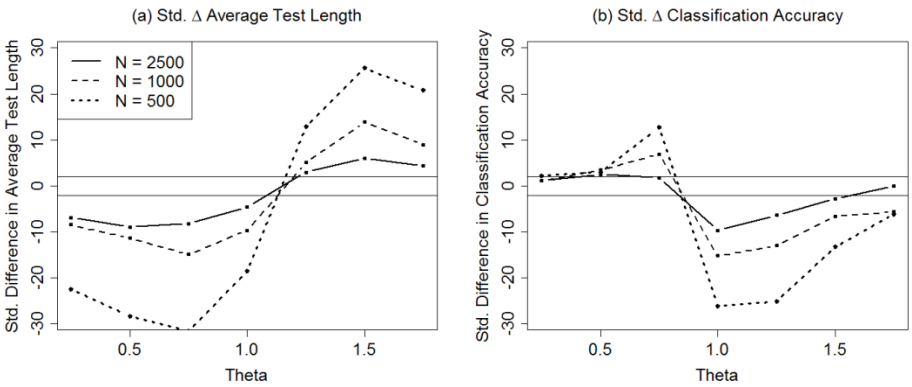

Figure 3:
Standardized Effect of *N* on Test Outcomes, Relative to Baseline Condition (40 Items, No Exposure Control)

We also hypothesized that small values of *N* would yield worse classifications at a fixed value of θ. Again for the condition with a maximum of 40 items and no exposure control, the top-left portion of Table 4 displays the classification accuracy (CA) for each value of *N*, as well as the difference between the *N* = 500 results and the baseline results. For θ values far from the cut, nearly all examinees were correctly classified; thus results for these examinees are not shown. But for examinees near the cut, small *N* yielded worse classifications above the cut but more accurate classifications below the cut. To confirm that these differences were not simply due to sampling error, we computed standardized differences in CA for each value of *N* relative to the baseline condition; these results are shown in Figure 3b. For $|\theta - \theta_0| \leq 0.5$, the standardized differences are again quite large. Under *N* = 500, for example, absolute standardized differences range from 3 to 26 units.

Thus we see two unexpected trends. First, when comparing results across different values of *N*, ATL and CA are inversely related. Second, the effects of *N* depend on whether θ is above or below the cut. To understand these trends, we examined the properties of the items selected under each sample size condition.

First, we determined whether the item selection criterion capitalized on spuriously large *a* estimates. For each possible test length *m* (*m* = 5, 6, 7,…, 40), Figure 4a displays the average discrimination parameter estimate for the *m* items comprising the test. And Figure 4b displays the average true *a* value for the corresponding items at each test length. (Notice that the curves for true and "estimated" parameters are identical in the $N = \infty$ condition.) As expected, the average $\hat{a}$ is inversely related to *N* at a fixed test length (Figure 4a). However, those items chosen under small *N* are actually less discriminating

Table 4:
Conditional Classification Accuracy ($\theta_0 = 1$)

		No Exposure Control ($d = 1$)					**Exposure Control ($d = 5$)**				
		θ Level					**θ Level**				
Max Length	N	0.50	0.75	1.00	1.25	1.50	0.50	0.75	1.00	1.25	1.50
40 Items	∞	99.1	89.0	50.1	88.6	99.2	97.5	84.2	49.4	83.6	97.7
	2500	99.3	89.4	46.8	87.3	99.0	97.8	84.3	49.7	83.4	97.3
	1000	99.3	90.5	45.0	86.0	98.7	97.9	84.7	47.6	82.6	97.3
	500	99.3	91.6	41.3	83.5	98.3	97.9	85.8	45.9	81.5	96.7
	Difference	**0.2**	**2.6**	**−8.9**	**−5.0**	**−0.9**	**0.4**	**1.7**	**−3.5**	**−2.1**	**−1.0**
80 Items	∞	99.7	93.4	49.4	93.6	99.8	99.0	87.9	50.4	88.1	99.0
	2500	99.8	94.1	47.3	92.4	99.7	98.9	88.4	50.1	87.9	98.8
	1000	99.8	94.3	45.6	91.4	99.6	99.2	88.7	48.8	87.3	98.8
	500	99.7	95.4	40.0	88.8	99.5	99.1	89.4	46.9	85.4	98.5
	Difference	**0.0**	**2.0**	**−9.4**	**−4.8**	**−0.4**	**0.1**	**1.4**	**−3.5**	**−2.6**	**−0.4**

than those chosen under large N (Figure 4b). When only item parameter estimates are available, some items may look very attractive in a statistical sense, but these items may not be truly optimal. If the average $\hat{a}$ is larger for small N at every test length, and larger a estimates should yield shorter tests, then why are tests not uniformly short for small N in the top-left portion of Table 3? To explain this discrepancy, we also investigated the relationship between true and estimated item difficulty. For each possible test length, Figure 5 displays the difference between the average b estimate and the average true b value for the corresponding items. Regardless of N, the selected items appear to be easier than they really are. That is, the items are more difficult than their difficulty estimates suggest, and this effect increases as N decreases.

How does this capitalization on underestimated $\hat{b}$ values explain the effects of N? By employing cut-based item selection, we expect examinees above the cut to have more correct responses than incorrect responses. But when N is small, these examinees are given items that are, on average, more difficult than they appear, resulting in more incorrect responses than we would expect. Thus these examinees require longer tests and tend to be misclassified more often. For those examinees below the cut, we already expect them to have more incorrect responses than correct ones. By administering items that are more difficult than they appear, the ratio of incorrect to correct responses increases, resulting in shorter tests and more accurate classifications. Thus the capitalization on $\hat{b}$ effectively moves the center of the indifference region toward the direction of true item difficulty.

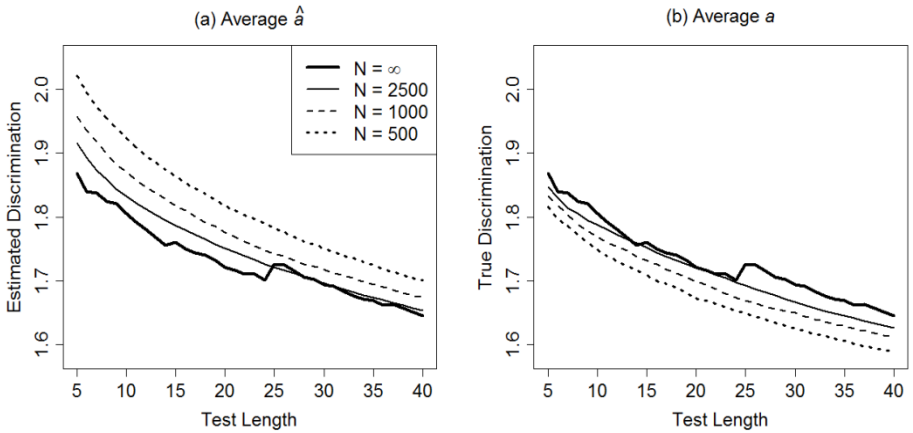

Figure 4:
Averages of Estimated and True Item Discrimination Parameters as a Function of Test Length (40 Items, No Exposure Control)

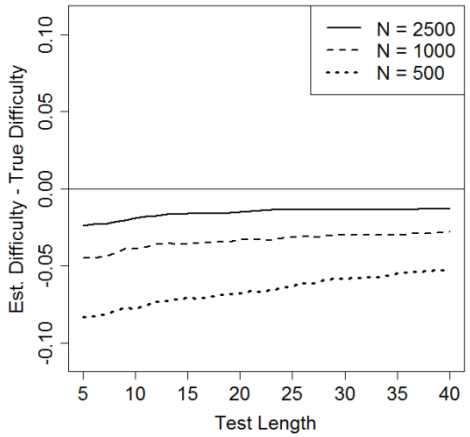

Figure 5:
Difference Between Estimated and True Item Difficulty Parameters as a Function of Test Length (40 Items, No Exposure Control)

Effects of exposure control and maximum test length

Returning to Table 3, notice that it contains ATLs (conditional on θ and N) for each combination of bin depth (d = 1 or 5) and maximum test length (40 or 80 items). First, regardless of the maximum test length, exposure control mitigates the effect of N by reducing capitalization on chance. This is consistent with our expectations. As the test proceeds, exposure control forces the item selection algorithm to reach further down the

list where items are in general less discriminating and we do not expect systematic differences between the estimated and true item parameter values. The increasing similarity between true and estimated item parameter values is apparent in Figures 4 and 5, but would be even more apparent if these figures were reproduced for the exposure control conditions. Second, what is the effect of allowing the test to run longer? Interestingly, the effect of *N* is much greater, regardless of whether exposure control is used. When only 40 items are allowed, many examinees reach the maximum test length. By allowing these examinees to continue testing, the differences among the values of *N* are exacerbated.

Returning to Table 4, we see that, as expected, exposure control mitigates the effect of *N* on CA, regardless of the maximum test length. Interestingly, although the longer maximum test length yields more accurate classifications for a given value of *N*, the effect of *N* (relative to the baseline condition) is similar to that when the maximum length is 40.

Results: Study 2

As we saw in the results from Study 1, the maximum KL information criterion tends to select items with *a* estimates containing positive calibration errors. But there is no reason to expect the *b* estimates of these items to be systematically underestimated in general. Rather, the sign of the discrepancy between estimated and true difficulty likely depends on a number of factors including the structure of the item pool and the cut location. To examine the effects of cut location, we conducted additional simulations with $\theta_0 = -1$ and $\theta_0 = 0.5$, each with a maximum of 40 items and no exposure control. For the original cut of $\theta_0 = 1$, Figures 6a and 6b display the differences in ATL and CA, respectively, for each value of *N* relative to the baseline condition. Figures 6c and 6d display results for the cut at $\theta_0 = 0.5$, and Figures 6e and 6f display results for the cut at $\theta_0 = -1$.

Interestingly, the effects of *N* on both ATL and CA are quite small when $\theta_0 = 0.5$. But when $\theta_0 = -1$, the effects of *N* are reversed relative to when $\theta_0 = 1$. In particular, tests are spuriously long for small *N* below the cut and spuriously short above it. This pattern can be explained using an argument similar to that given in the previous section. Namely, when $\theta_0 = -1$, examinees are administered items with spuriously high *b* estimates, making items appear more difficult than they really are.

Discussion

Our goal was to examine the effects of capitalization on chance on test length and classification accuracy in variable-length tests employing the SPRT and cut-based item selection. Contrary to our expectations, capitalization on large discrimination estimates did not always yield shorter tests or worse classifications. The reason for these unanticipated findings was capitalization on spuriously high or low difficulty estimates. In addition, the effects of capitalization on *b* estimates were highly sensitive to the cut location and true value of the latent trait. We also found that implementing exposure control reduced the effects of capitalization on chance, whereas allowing for a longer maximum test length appeared to increase the effects.

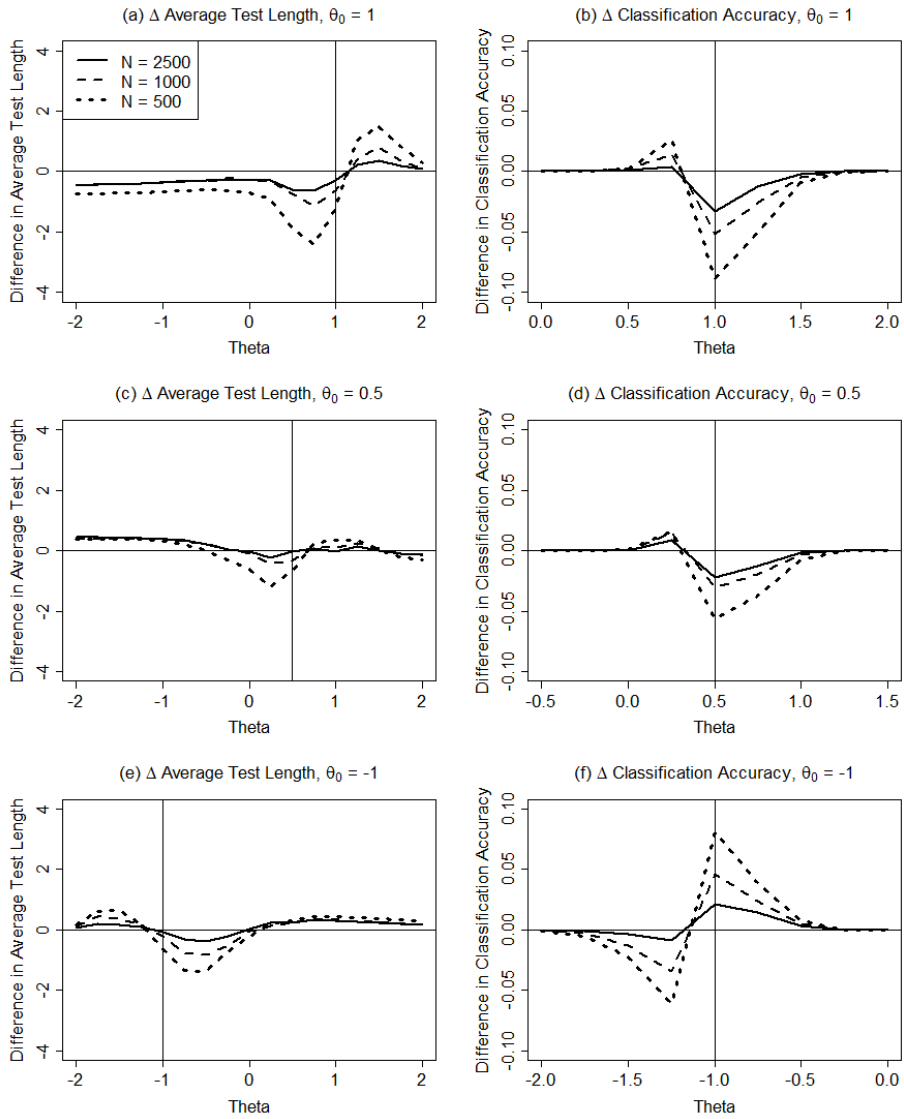

Figure 6:
Effect of Cut Location on Test Outcomes (40 Items, No Exposure Control)

In contrast with the results of Spray and Reckase (1987) and Kalohn and Spray (1999), we have demonstrated that treating item parameter estimates as the true parameter values can have practically important effects on both test length and classification accuracy. The reason for the conflicting results is in the simulation design. By computing the dependent

measures conditional on several true values of the latent trait, we discovered, for example, that capitalization on chance increased test lengths on one side of the cut but decreased test lengths on the other side. The previous studies averaged dependent measures over an entire sample of examinees, possibly obscuring these trends. Additionally, these studies employed item parameter estimates based on a single, large pretest sample. By manipulating the size of the pretest sample, we were able to observe the effects of various magnitudes of item calibration error.

Previous research on adaptive item selection has demonstrated that several popular item selection criteria tend to choose items with spuriously high discrimination estimates (van der Linden & Glas, 2000). Classification tests, on the other hand, often seek to match items with the cut point (Eggen, 1999). In this study, we have demonstrated that cut-based item selection may capitalize not only on spuriously high *a* estimates, but also on spuriously high or low *b* estimates. However, our results suggest that whether the item selection criterion favors over- or underestimated *b* values is difficult to predict. It likely depends on many factors including the structure of the item pool, the cut location, and the true value of the latent trait.

Another unexpected finding was that test length and classification accuracy were inversely related across the values of *N*. Our hypothesis was that small *N* would yield shorter tests, and if examinees were given fewer items, classification accuracy would also suffer. For a given value of *N*, this is exactly what would happen: if we administer 40 items instead of 20 (holding everything else constant), repeated administrations would show more accurate classifications for the longer test. But under the SPRT, the test ends when there is sufficient evidence to make a decision. So in the condition with a maximum of 40 items and no exposure control, examinees above the cut were administered more items under small *N* than in the baseline condition precisely because the examinees under small *N* were more difficult to classify. But we cannot expect the SPRT to perform well when *N* is small because the test statistic is evaluated with respect to item parameter estimates that may be far from their true values. Thus, even if tests are longer under small *N*, classification decisions will not necessarily be more accurate. As mentioned in the introduction, examinees in the indifference region tend to have longer tests and be misclassified more often than examinees outside of it. So it appears that the effect of capitalization on difficulty estimates is to move the indifference region in the direction of true item difficulty, though further research is needed to better understand this phenomenon.

However, we note that even in the baseline condition, the SPRT exceeded the nominal Type I and Type II error rates for examinees at the bounds of the indifference region. This occurred simply because the test was not allowed to run long enough for these examinees. This is clear in Table 4: at $\theta = 0.75$ and $\theta = 1.25$, classification accuracy noticeably increases when the maximum test length is increased to 80 items, regardless of whether exposure control is used.

This study can be extended in several ways. First, in addition to exposure control, most operational testing programs impose other non-statistical constraints on item selection such as content balancing. These constraints reduce the dependence of item selection on

statistical criteria, thus reducing capitalization on chance. Second, our simulations employed only one formulation of the SPRT; recently, other variations have been proposed. One variation is the generalized likelihood ratio test (Thompson & Ro, 2011) in which the numerator or denominator of the likelihood ratio is evaluated at the maximum of the current likelihood function for θ_1. If the maximum is located outside the indifference region, the generalized ratio will be more extreme than that evaluated at the endpoints of the indifference region, resulting in shorter tests. Another variation is the stochastically curtailed SPRT (Finkelman, 2008). After each response, one computes the probability of changing classifications from the current one. If this probability is sufficiently small, the test ends. In this formulation, the effects of calibration error depend on the relationship between item parameters and computation of this probability.

A third factor which might moderate the effects of capitalization on chance is the choice of IRT model. For a fixed pretest sample size, 3PLM parameters tend to be less precisely estimated than those of the 2PLM, due to difficulties in estimating the pseudo-guessing parameter (Thissen & Wainer, 1982). Thus under the 3PLM, one might expect greater effects on test outcomes than those reported in this study. Under the 1PLM, however, items are allowed to vary only with respect to location. For fixed *N*, these estimates tend to be more precise than those of more complex models, and item selection would consider only the proximity of items to the cut. Thus the effects of capitalization on chance might be reduced relative to those reported in this study.

Funding statement

This research was supported by a 2010 CTB/McGraw-Hill Innovation Research and Development grant.

References

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

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

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

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

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

Lin, C.-J., & Spray, J. (2000). *Effects of item-selection criteria on classification testing with the sequential probability ratio test* (Research Report 2000-8). Iowa City, IA: ACT, Inc.

R Development Core Team (2011). R: A language and environment for statistical computing (Version 2.13.2) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Available from 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 theory and computerized adaptive testing* (pp. 237-254). New York: Academic Press.

Spray, J. A., & Reckase, M. D. (1987). *The effect of item parameter estimation error on decisions made using the sequential probability ratio test* (Research Report ONR 87-1). Iowa City, IA: ACT, Inc.

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.

Thissen, D. & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*, 397-412.

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

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

Thompson, N. A., & Ro, S. (2011, April). *Likelihood ratio based computerized classification testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

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

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